Supplementary Material: Lite2Relight

ACM Reference Format:
. 2024. Supplementary Material: Lite2Relight. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3641519.3657470

Code and pre-trained checkpoints is available under https://vcai.mpi-inf.mpg.de/projects/Lite2Relight//.

1 IMPLEMENTATION DETAILS

1.1 Training and Testing Details

We implemented our method using PyTorch [Paszke et al. 2019]. $\mathcal{R}$ comprises 14 MLP layers with ReLU activations. It is designed to optimize the objective functions detailed in Sec 3.4. For optimization, we employ the Adam optimizer [Kingma and Ba 2015] with a learning rate of 0.0003. The weights for the losses $\mathcal{L}_{\text{lat}}$, $\mathcal{L}_{\text{C}}$, and $\mathcal{L}_{\text{LPIPS}}$ are set at $\lambda_0 = 10, \lambda_1 = 0.01, \lambda_2 = 1.0$, respectively.

We utilize the pretrained checkpoints of EG3D, $\mathcal{E}$, and *AFA* from their original implementations. Our training dataset includes images of 250 subjects captured from two frontal viewpoints, each subject being relit under 50 different natural illumination conditions.

The entire training process is completed in approximately 16 hours over two A100/A40 GPUs. This is achieved across 27k iterations with a batch size of 8.

In a testing scenario with a monocular image, our approach requires around 140*ms* for inversion and relighting. Once the subject is embedded in the target illumination environment, we can render novel viewpoints at a rate of 31 frames per second on an NVIDIA A40 GPU.

1.2 Interactive Demo Details

In addition to our technical evaluations, we showcase the practical effectiveness of Lite2Relight with a user demo, powered by a web camera on a workstation equipped with a single NVIDIA 3090 GPU. This live demonstration involves real-time tracking and alignment of the user's head using facial landmarks. For each frame, our method efficiently inverts the subject, applies the selected environment map, and performs relighting and viewpoint rendering dynamically, as illustrated in Fig. 1. Currently, our demo operates at a rate of 7 frames per second. We wish to highlight that with further engineering and optimization, there is significant potential to enhance the runtime performance of our demonstration.

Additionally, we provide a comprehensive video showcasing the capabilities of our user demo in the supplementary materials. This video aims to give viewers a more tangible sense of the real-time interactivity and effectiveness of Lite2Relight in a live setting.

2 ABLATION STUDY

Additional Evaluation Metrics: To further emphasize on quality of relighting, we provide additional evaluation metrics such as LPIPS [Johnson et al. 2016], RMSE and DISTS [Ding et al. 2020] loss metrics. We observe that Lite2Relight convincigly outperfroms the baseline methods as shown in Tab. 1.

Table 1: Quantitative Results: Ablation Study: Additional Metrics. We report LPIPS, RMSE and DISTS metrics in addition to SSIM, landmarks distance (LD), and PSNR on the test data of lightstage, where subjects are relit under novel viewpoints.

	LPIPS↓	RMSE↓	DISTS↓
PhotoaApp	0.4163	0.1988	0.2031
NeRFFaceLighting	0.2966	0.2905	0.2409
Lite2Relight	**0.2493**	**0.1841**	**0.1718**

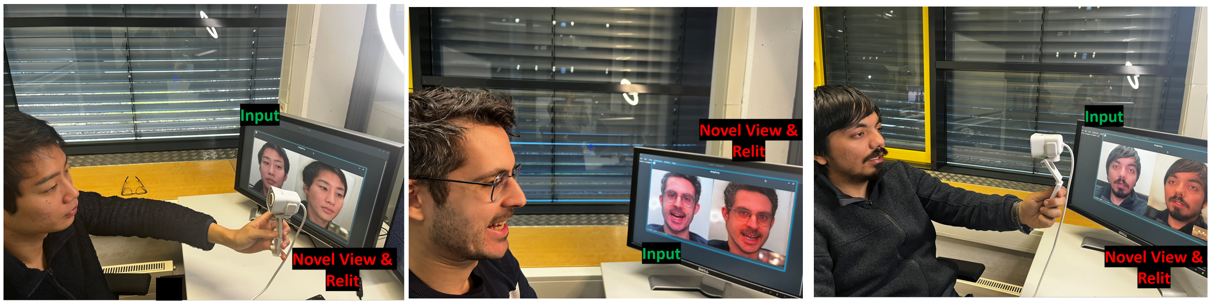
Number of Viewpoints: As our 3D face prior was derived from monocular data, we delved into examining the necessity of multiple viewpoints in the training of our relighting network, $\mathcal{R}$. To this end, we trained various iterations of $\mathcal{R}$ using different sets of viewpoints—specifically, 1, 2, 4, and 8 viewpoints. The outcomes of these experiments are systematically presented in Tab. 2. A noteworthy observation from our study is the robustness of our approach to the number of training viewpoints, as evidenced by the consistently high PSNR and SSIM metrics across different variants. This claim is further substantiated by our qualitative results in Fig. 2, where negligible differences are observed in the renderings, even for extreme profile views, as indicated in rows 1 and 3.

Taking into account the Landmark Distance (LD) scores, we identified that utilizing two frontal viewpoints represents the optimal training configuration for $\mathcal{R}$. This decision stems from the fact that frontal viewpoints comprehensively cover most facial regions, allowing $\mathcal{E}$ to accurately invert these views while ensuring minimal identity loss.

Table 2: Quantitative Results: Ablation Study: Number of Viewpoints. We report SSIM, landmarks distance (LD), and PSNR on the test data of lightstage, where subjects are relit under novel viewpoints.

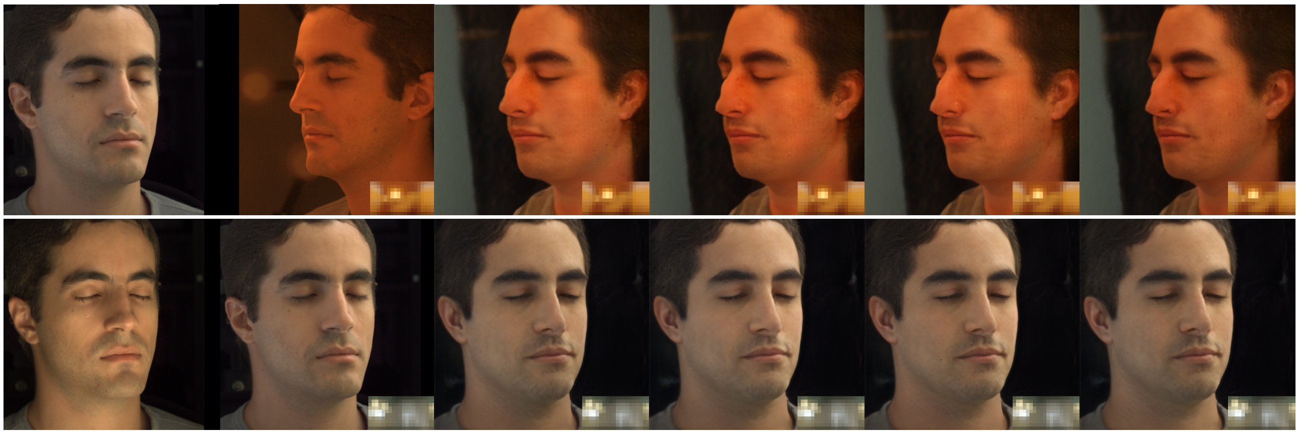
	SSIM ↑	LD ↓	PSNR ↑
Views = 1	0.831	10.45	28.31
Views = 4	0.834	10.26	28.30
Views = 8	0.834	10.1	28.31
Views = 2	**0.834**	**9.76**	**28.33**

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657470



Subject 1 Subject 2 Subject 3

Figure 1: Interactive Demo: We show that Lite2Relight can be driven from webcam and consumer hardware at interactive rates.



Input GT 8 views 4 views 2 views 1 view

Figure 2: Ablation Study: Number of Viewpoints. first column: Input Image, second column: groundtruth for a novel view. Subsequent columns: prediction for the same novel view, trained under multiple training views. Our approach is robust to the number of training views and performs well when trained with even monocular inputs.

Number of Subjects: We conducted an ablation study to evaluate the impact of training subject quantity on generalization performance, aiming to emphasize the benefits of integrating generative priors with supervised learning approaches. Our lightstage dataset comprises 353 subjects, and for this study, we conducted experiments with subsets of 250, 50, and 10 subjects. Our quantitative analysis revealed that the model trained with 250 subjects achieves the best performance. However, it is noteworthy that the difference in performance between this model and the one trained with as few as 10 subjects is relatively marginal. This observation, indicative of robust generalization, is further substantiated in Fig. 3. Here, we demonstrate the ability of our method to perform 3D consistent view and illumination editing on an unseen subject under various lighting conditions, as seen in rows 2 and 3.

Our observations in Sec. 2 and Sec. 2, suggest that Lite2Relight does not necessitate a densely-equipped multiview lightstage setup or extensive data collection campaigns. This advantage significantly reduces the complexities associated with hardware and data storage, thereby paving the way for more feasible and generalizable portrait relighting solutions. Such solutions could potentially be realized with minimal equipment, akin to the approach proposed by Sengupta et al.[Sengupta et al. 2021], which utilizes a few desktop monitors. However, the exploration of this avenue falls outside the scope of our current project and we leave this for future work.

Table 3: Quantitative Results: Ablation Study: Number of Subjects. We report SSIM, landmarks distance (LD), and PSNR on the test data of lightstage, where subjects are relit under novel viewpoints.

	SSIM ↑	LD ↓	PSNR ↑
Subjects = 10	0.829	10.24	28.26
Subjects = 50	0.829	9.61	28.31
Subjects = 250	**0.834**	**9.76**	**28.33**

Latent Editing: The utilization of a feedforward encoder-based inversion in our framework extends beyond merely accelerating inference. A significant advantage of this approach is that the encoded latent vector reliably remains within the latent manifold of the generator. This ablation study aims to merge the semantic manipulation capabilities inherent in the latent space with the task of relighting. It shows that our $\mathcal{R}$ effectively operates within the rich

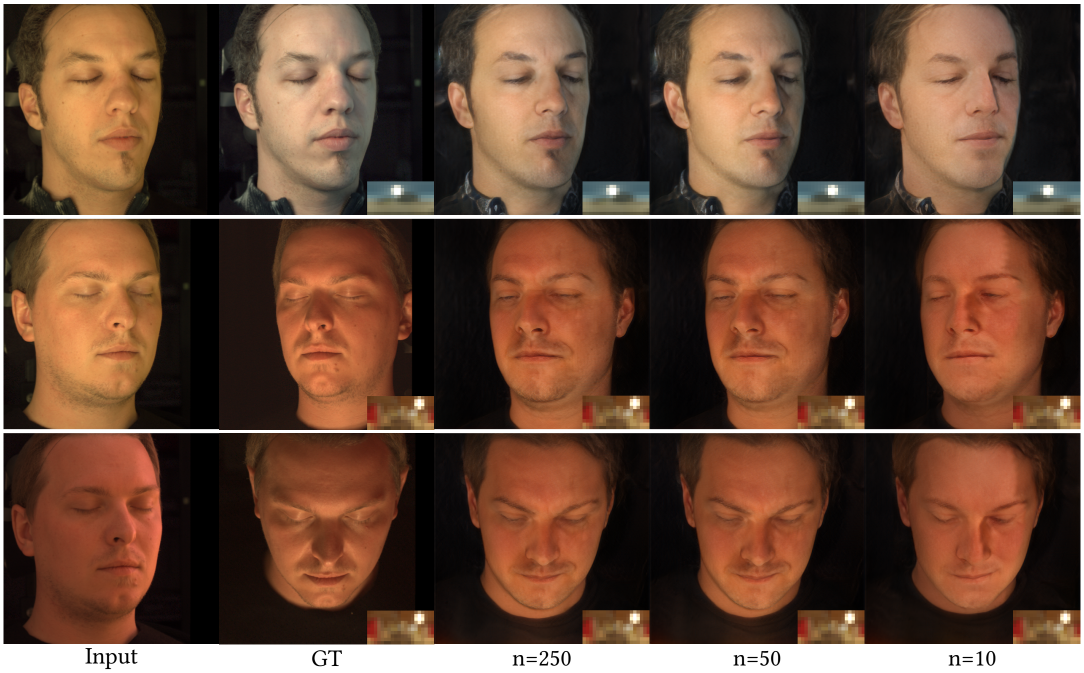


Figure 3: Ablation Study: Number of Subjects. first column: Input Image, second column: groundtruth for a novel view. Subsequent columns: prediction for the same novel view, trained with different number of subjects. Results indicate that we have better results with more training subjects.

latent manifold of EG3D. Leveraging the latent attribute directions identified by GOAE [Yuan et al. 2023], we demonstrate the simultaneous editing of viewpoint and illumination alongside changes in specific facial attributes—namely, “age”, “anger”, and “glasses”. These multifaceted edits are showcased in Fig. 4, illustrating the robust and versatile nature of our method.

3 CHALLENGES AND FUTURE WORK

While our method demonstrates effective photorealistic editing of viewpoints and relighting, there remain areas for improvement.

Occlusions and non-frontal views: Our approach depends on the pretrained 3D-aware encoder for inversion, which positions the given portrait within the canonical 3D space of the generator. However, this encoder was primarily trained on unoccluded, front-facing views. As a result, the performance of the encoder, and consequently our relighting technique, encounter challenges with non-frontal views, the presence of accessories, and occluded faces as shown in the Fig. 5

Hard Shadows: To achieve relighting at interactive rates, Lite2Relight relies on relighting within the latent manifold of the generator as opposed to explicit modelling face reflectance through HDR OLAT rendering as in [Rao et al. 2023] or explicit specular and diffuse maps as in Total Relighting [Pandey et al. 2021]. Consequently currently casting hard shadows is a challenge.

OLAT Synthesis: Furthermore, even though we use a lightstage dataset, currently, our approach does not support the synthesis of HDR (High Dynamic Range) OLAT images, restricting our ability to recreate unconventional lighting conditions. This limitation stems from the fact that OLAT images do not fall within the distribution of the pretrained generator. While training an OLAT-based generator capable of predicting a dense reflectance basis might address this issue, it lies beyond the scope of our current project. Nonetheless, exploring such a generator presents a fascinating and potentially impactful avenue for future research, promising to further advance the field of photorealistic relighting and editing.

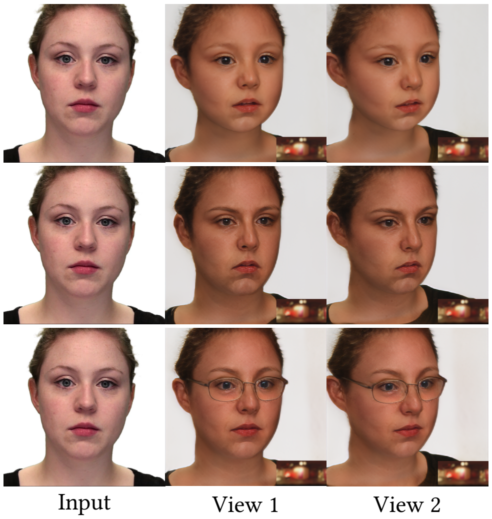


Figure 4: Ablation Study: Latent Editing First column have the input. The remaining columns have the respective subjects manipulated by the latent attributes following [Yuan et al. 2023].

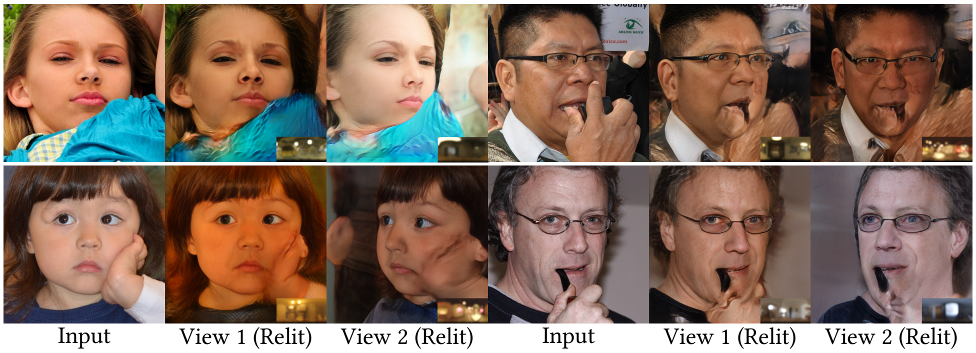


Figure 5: Challenges: Our method struggles to modify viewpoint and illumination accurately in the presence of extreme head poses or occlusions

REFERENCES

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. 2020. Image Quality Assessment: Unifying Structure and Texture Similarity. *CoRR* abs/2004.07728 (2020). https://arxiv.org/abs/2004.07728

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 694–711.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. San Diega, CA, USA.

Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* (2021).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pramod Rao, Mallikarjun B. R, Gereon Fox, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Fangneng Zhan, Ayush Tewari, Christian Theobalt, and Elgharib Mohamed. 2023. A Deeper Analysis of Volumetric Relightable Faces. *International Journal of Computer Vision* (10 2023), 1–19. https://doi.org/10.1007/s11263-023-01899-3

Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. 2021. A Light Stage on Every Desk. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 2400–2409.

Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. 2023. Make Encoder Great Again in 3D GAN Inversion through Geometry and Occlusion-Aware Encoding. *arXiv preprint arXiv:2303.12326* (2023).