Supplemental Materials: Personalized Video Relighting With an At-Home Light Stage

Jun Myeong Choi, Max Christman, and Roni Sengupta

University of North Carolina at Chapel Hill, NC 27514, USA
{chedgekr,mrlc,ronisen}@cs.unc.edu

A Overview of Appendices

Our appendices contain
Temporal Consistency (Sec. B)

- In Fig. 10, we compare the temporal changes in skip-connected features between our method and the approaches by [1,2]. Additionally, we demonstrate that the use of the LCFN module contributes to improved temporal consistency, showing smaller differences in feature space. (Sec. B.1)
- We provide Fig. 1 in video **'main.mov'** (Sec. B.2). We also demonstrate the result of rotating the panorama in Fig. 11.
- We present video **'compare.mov'** demonstrating that our method exhibits temporal consistency compared to the approaches of [1,2] (Sec. B.3).

Others (Sec. C)

- We provide implementation details of our network architecture. (Sec. C.1)
- We perform ablation studies evaluating the effects of data pre-processing. (Sec. C.2)

Image Comparison (Sec. D)

- In Fig. 12, we demonstrate our method is robust to identity transformation and facial decorations.
- Additional examples of visual comparisons shown in Fig. 4, 5 are demonstrated with the LSYD dataset in Fig. 13, and the OLAT dataset in Fig. 14.
- In Tab. 3, we present a quantitative comparison illustrating the performance improvements brought about by our LCFN and monitor prediction modules. The supporting visual results can be observed in Fig. 15.
- Additional instances of the visual comparison in real-world scenarios, as depicted in Fig. 6, are displayed in Fig. 16.

B Temporal Consistency

B.1 Feature Difference

In Fig. 10, as evident, the sharp spikes in values occur when L_{src} undergoes significant changes. In such cases, the difference in skip-connected features is larger compared to scenarios where L_{src} undergoes small variations. These substantial differences in feature values contribute to temporal inconsistency. Regardless of the magnitude of the change in L_{src}, our objective is to transmit only the shape and characteristics of the portrait through skip-connected features, without the information about the reflection of light on the face. Therefore, to minimize this difference, we demonstrate that by employing light conditioned feature normalization (LCFN) and de-lighting the features, we can enhance temporal consistency as shown in red line.

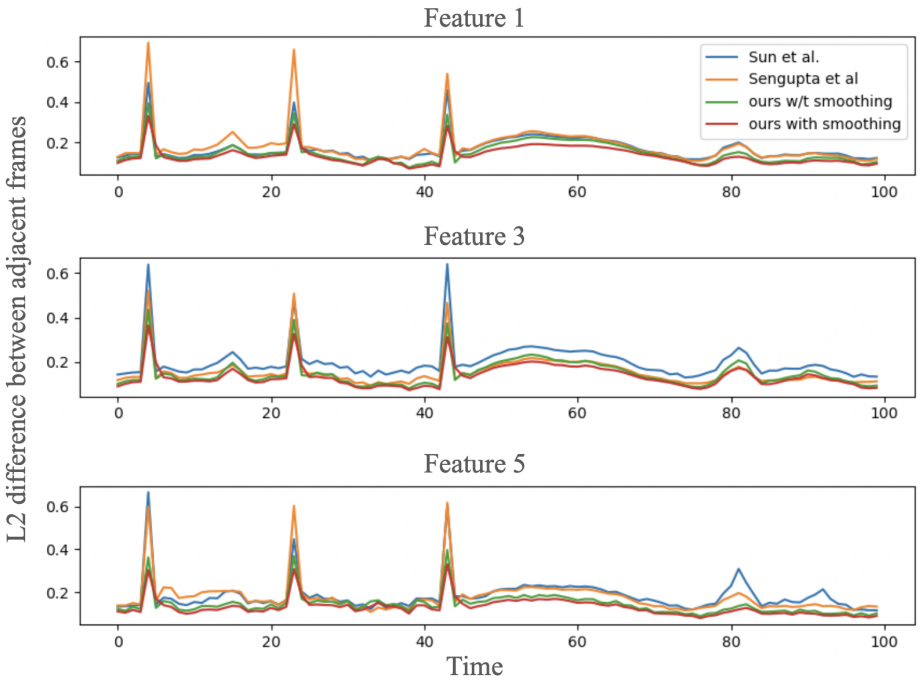

Fig. 10. Plots illustrate the L2 distance of skip-connected features (Feature 1, 3, and 5, respectively) between adjacent frames, relighting 100 consecutive portrait images into a single target light. The blue line corresponds to Sun *et al.* [1], the orange line to Sengupta *et al.* [1], the green line to our method without LCFN, and the red line to our method with LCFN.

Fig. 11. We relight an input image with a rotating panorama. Red box indicates the portion of the panorama projected as monitor

B.2 Relit Video

In the **'main.mov'** video, we demonstrate how we captured the Light Stage at Your Desk (LSYD) data and showcase the relit results, as described in Fig. 1. From 0 to 7 seconds, we illustrate the process of capturing. The video on the left shows the capturing procedure, with the top-right corner displaying the portrait video and the bottom featuring the corresponding monitor video. From 7 to 42 seconds, we present the relighting results. On the left side, the input portrait and monitor are displayed, while on the right side, the relit portrait and the target monitor are shown. (Relighting results for different target monitors are presented approximately every 7 seconds.) Additionally, in Fig. 11, we also relight a face with a rotating panorama.

B.3 Relit Video Comparison

In the **'compare.mov'** video, we conduct a comparison with Sun *et al.* [2] and Sengupta *et al.* [1] in terms of temporal consistency, utilizing the ideal ring light as the target light. This video supports two playback speeds (In addition to the original speed video, we provide a 0.5x slowed-down relit video to facilitate a clearer observation of temporal consistency). In the top-left corner is the input portrait, and in the top-right corner is the relit result of ours with LCFN and L_{src_avg}. In the bottom-left corner is Sengupta *et al.* [1]'s relighting result, and in the bottom-right corner is Sun *et al.* [2]'s result. We note that our results are more temporally consistent.

C Others

C.1 Implementation Details

Each portrait image has a resolution of 480×480, and each monitor image is 18×32. The input images are cropped to the subject's head to limit the effect of the background. For the *light decoder*, we use convolutional layers to maintain a resolution of 30x30, while changing the channel size from $448 = 256+128+64$ to $2304 = 4 \times 18 \times 32$. Then, we performed 30×30 average pooling to downsize the feature map to $4 \times 18 \times 32$. Finally, we did a weighted average to obtain a final feature map size of $3 \times 18 \times 32$. All convolutional and MLP layers are followed by pixel normalization and a PReLU activation function. We use $\lambda_{L1} = 1$, $\lambda_P = 0.1$, $\lambda_C = 0.5$, $\lambda_D = 0.1$ and $\lambda_G^M = 0.5$. We train the generator and discriminator with the Adam optimizer, with a learning rate of 10^{-3} and 10^{-6}, and a batch size of 2.

C.2 Data Pre-processing

In Sec. 3.1 we discussed how facial keypoint detection enables robustness to relighting with respect to pose and expression, unlike face parsing proposed in Sengupta *et al*. [1]. This improvement can be seen in Tab. 4.

Table 4. We observe that keypoint-based source-target data pairing improves upon previous [1] face parsing-based pairing methods.

	Ours		Sengupta *et al*. [1]	
	LPIPS ↓	**DISTS ↓**	**LPIPS ↓**	**DISTS ↓**
Segment	0.1116	0.1031	0.1229	0.1123
Keypoints	**0.0966**	**0.0932**	**0.1209**	**0.1110**

D Image Comparison

Robustness w.r.t. facial decorations. We observe that our method can generalize to various face decorations, e.g., glasses, earrings, and makeup, that are not present during training, see Fig. 12, col 2-4 for visuals, and row 4 for quantitative evaluation across 648 examples. Note that the performance deteriorates slightly from the test set without any face decorations (RMSE: 8.21, LPIPS: 0.08). We believe adding examples of face decoration during training will further improve the performance, e.g., Sengupta *et al*. [1] showed relighting on glasses by training on them. We can showcase individuals with different hair colors in the final version.

Stability w.r.t. identity transformation. Our method is robust to identity transformation, yielding an average RMSE 6.08 and LPIPS: 0.06 (Fig. 12, col-5).

Fig. 12. We capture additional test data to show robustness w.r.t. unseen pose, expression, glasses, piercing, and makeup during testing, and report an average RMSE and LPIPS error across 648 images. Our method is robust w.r.t identity transformation.

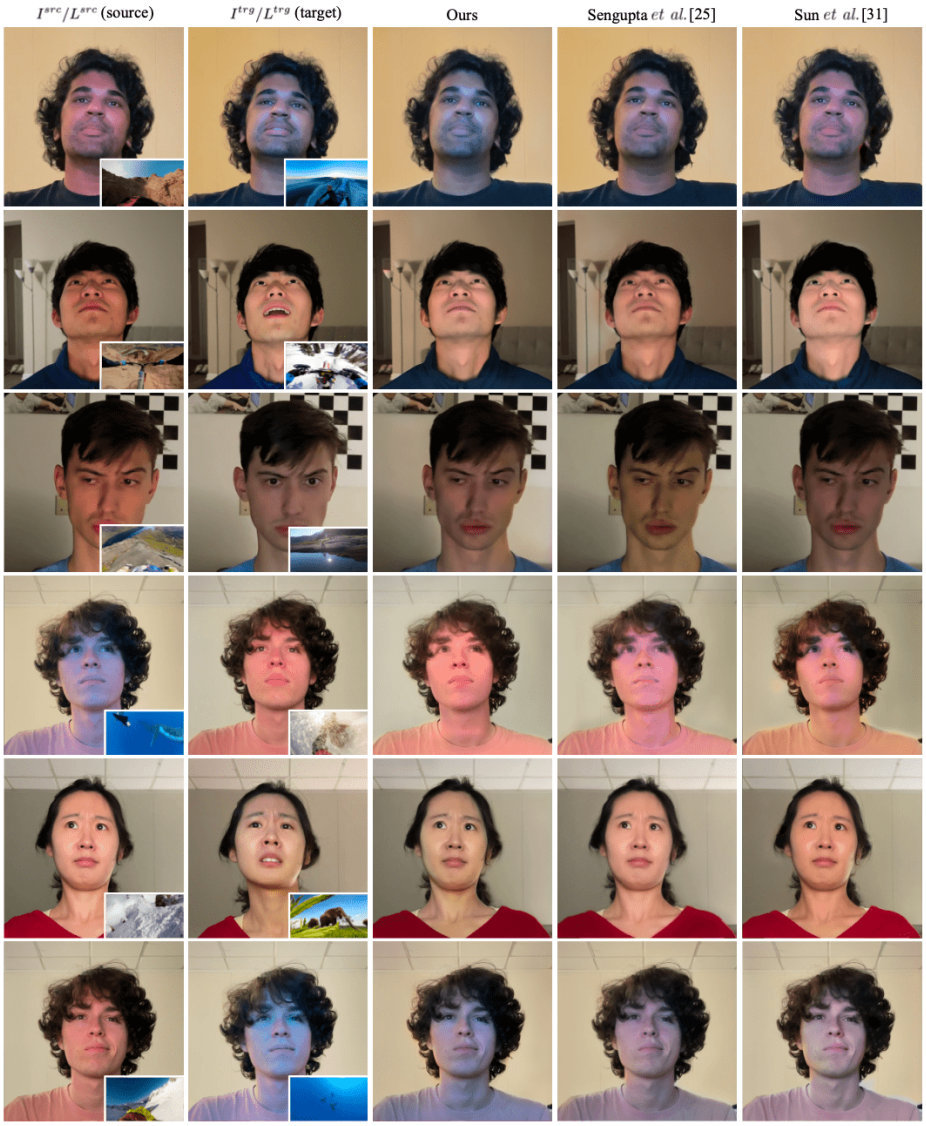

Fig. 13. In addition to Fig. 4, we conduct a visual comparison with established relighting techniques [1,2] using unseen test LSYD data. Our approach (Col. 3) yields notably superior relighting outcomes in contrast to existing methods (Cols 4 and 5).

Fig. 14. In addition to Fig. 4, we perform a qualitative comparison with established relighting techniques [1,2] using unseen test OLAT data [3]. Our method (Col. 3) produces better relighting results compared to existing approaches (Cols 4 and 5).

Fig. 15. We present visual evidence supporting the observations outlined in Tab. 3. In comparison with Col. 4 (without the LCFN module), 5 (without L_{src} prediction), and 6 (without both the LCFN module and L_{src} prediction), we note that the proposed modules LCFN and L_{src} prediction exhibit substantial enhancements in our result (Col. 3).

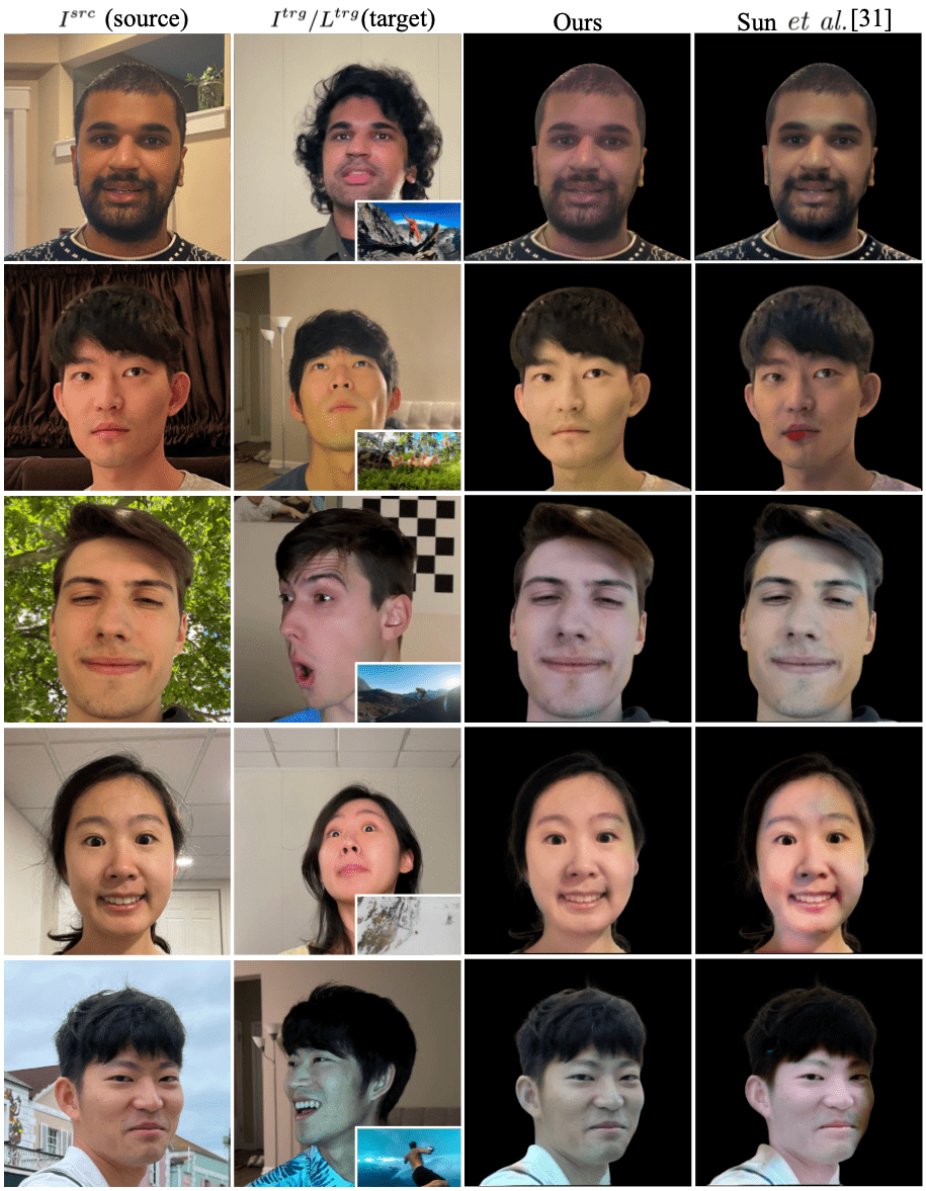

Fig. 16. Additional results for Fig. 6

References

1. Sengupta, S., Curless, B., Kemelmacher-Shlizerman, I., Seitz, S.M.: A light stage on every desk. CoRR **abs/2105.08051** (2021), https://arxiv.org/abs/2105.08051
2. Sun, T., Barron, J.T., Tsai, Y., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P.E., Ramamoorthi, R.: Single image portrait relighting. CoRR **abs/1905.00824** (2019), http://arxiv.org/abs/1905.00824
3. Zhang, L., Zhang, Q., Wu, M., Yu, J., Xu, L.: Neural video portrait relighting in real-time via consistency modeling (2021)