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

Abstract. In this paper, we develop a personalized video relighting algorithm that produces high-quality and temporally consistent relit videos under any pose, expression, and lighting condition in real-time. Existing relighting algorithms typically rely either on publicly available synthetic data, which yields poor relighting results, or on actual light stage data which is difficult to acquire. We show that by just capturing recordings of a user watching YouTube videos on a monitor we can train a personalized algorithm capable of performing high-quality relighting under any condition. Our key contribution is a novel image-based neural relighting architecture that effectively separates the intrinsic appearance features - the geometry and reflectance of the face - from the source lighting and then combines them with the target lighting to generate a relit image. This neural architecture enables smoothing of intrinsic appearance features leading to temporally stable video relighting. Both qualitative and quantitative evaluations show that our architecture improves portrait image relighting quality and temporal consistency over state-of-the-art approaches on both casually captured 'Light Stage at Your Desk' (LSYD) and light-stage-captured 'One Light At a Time' (OLAT) datasets. Source code is available at `https://github.com/chedgekorea/relighting`

Keywords: Image-based Relighting · 'At Home' Light Stage

1 Introduction

With the recent rise in popularity of video conferencing for business, educational, and personal activities, there is a significant demand for improving facial lighting. Virtually relighting our images and videos helps us improve the appearance of our faces without requiring explicit studio-quality lighting in a dedicated space or any specialized lighting expertise. Recent advances in deep neural networks have renewed interest in the problem of virtual relighting.

Training a deep neural network for relighting requires extensive training data that includes source images paired with relit target images. One way of acquiring this data is by using a large spherical rig with numerous lights and cameras, known as a light stage [5]. While light stage data has been shown to produce high-quality relighting results [36,44,23,38,22,48], the limited availability of datasets, trained models, and access to the light stage itself has impeded further research.

(a) LSYD   (b) Input Image   (c) Relit-Net   (d) Output

Fig. 1. We learn a personalized relighting algorithm that generates temporally consistent and high-quality portrait videos under different lighting. We train the network using recordings of users watching YouTube on a monitor, thereby creating a Light Stage at Your Desk (LSYD). We project a portion of the LDR environment map with a 180 ° FoV as the monitor light, while a portion of the remaining 90 ° FoV is mapped as the background. We can achieve a harmonization effect with the virtual background.

For example, *One Light At a Time* (Dynamic-OLAT) [44] is the only publicly available light stage relighting dataset consisting of four individuals only. As a result, researchers have often turned to synthetic data to train their relighting algorithms [31,49,13,35]. Unfortunately, existing synthetic data compromises the quality of relit images.

We draw inspiration from recent work by Sengupta *et al.* [30,26] and develop a personalized relighting model by capturing a single user's appearance while being lit by a computer monitor. However, the results and applications of Sengupta *et al.* [30] have several limitations. First, it requires capturing users with fixed poses and expressions, an unrealistic requirement for actual users that impedes casual capture. Second, the relighting algorithm requires a dark room with negligible ambient light, limiting the environment in which this can be applied. Third, the resulting relit video is temporally unstable, exhibiting significant flickering artifacts, making it unsuitable for relighting Zoom calls. Lastly, it requires knowledge of source lighting and thus is unable to relight arbitrary portrait images captured in the wild.

In this paper, we show that casually captured light stage data is sufficient to develop a high-quality temporally consistent video portrait relighting algorithm that works under arbitrary conditions (i.e. pose, expression, and ambient lighting) in real-time (∼45 fps). To that end, we create our own casually captured light stage dataset with varying pose, expression, and lighting, called *Light Stage at Your Desk* (LSYD). Our key contribution is a neural image-based relighting architecture, based on the commonly used U-Net [36,49,22,40,38], that better disentangles the source lighting from the user's intrinsic facial appearance (shape and reflectance) and then adds back the target lighting to generate a relit image. Existing image-based relighting architectures [36,30] fail to accurately separate source lighting information from intrinsic appearance features in the encoder, leading to inconsistent and temporally unstable video relighting. To this end, we introduce the *light-conditioned feature normalization* (LCFN) module, which performs relighting and also predicts the source lighting from an input image. The LCFN module also enables temporal stability by performing exponential smoothing of *de-lit* intrinsic appearance features and facilitates relighting
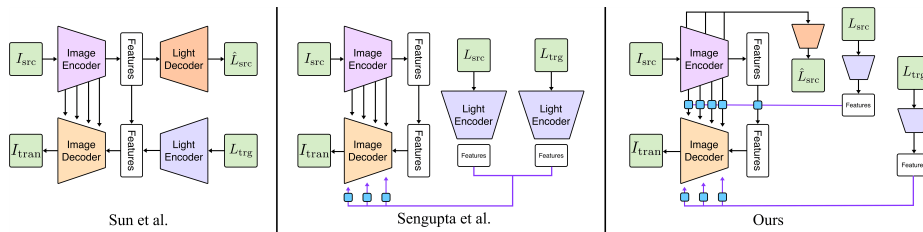
Fig. 2. We highlight the key structural differences between our relighting architecture and that of [36,30]. Our approach removes source lighting information from input image features and only propagates intrinsic appearance (geometry and reflectance) features from the encoder to decoder, which results in better relighting quality and more temporal stability. In contrast [36,30] propagates entire image features from the encoder to the decoder without 'de-light', and expects the decoder to remove source lighting and add target lighting information.

of any arbitrary portrait image with unknown source lighting. We also improve the data pre-processing pipeline from Sengupta *et al.* [30] to make the relighting algorithm more robust to pose, expression, and ambient lighting conditions.

We compare our relighting network with two other algorithms: Sun *et al.* [36], which was originally trained on light stage data (OLAT), and Sengupta *et al.* [30], which was originally trained on casually captured data (albeit with fixed pose, expression, and no ambient lighting). For a fair comparison, we train all algorithms for personalized relighting using the same data pre-processing steps and loss functions on 5 individuals from our LSYD dataset and 4 individuals from OLAT [44]. Our network outperforms Sun *et al.* [36] and Sengupta *et al.* [30] by 22.3% and 23.6% respectively on the LSYD dataset and by 23.5% and 25.6% on the OLAT dataset, in terms of LPIPS. Qualitatively our method produces superior relighting in terms of color and quality. We further show that our approach is more temporally consistent, leading to less flickering than Sun *et al.* [36] or Sengupta *et al.* [30]. Detailed ablation studies show that LCFN and source monitor prediction improves relighting quality, feature, and source monitor smoothing improves temporal consistency, and data pre-processing improves robustness to pose and expression.

To summarize, our main contributions are as follows:

• We show that casually captured *Light Stage at Your Desk* (LSYD) data can be used to build a high-quality temporally consistent personalized video relighting algorithm without requiring access to an expensive light stage setup.

• We introduce a novel video relighting architecture that separates the source lighting from the user's intrinsic appearance features and then adds back the target lighting, leading to improved relighting and temporal consistency for videos.

• While our relighting network focuses on 'at home' Light Stage (LSYD dataset) outperforming state-of-the-art algorithms, we also perform equally well on actual Light Stage captured OLAT [44] datasets, and any arbitrary portrait image captured 'in the wild'.

2   Related Work

Portrait relighting methods change the appearance of the face to match a target lighting condition. This can be expressed through lighting parameters (e.g., an environment map, spherical harmonics, directional lighting, etc.) or through a reference image of another person. Our approach relights a portrait image to a lighting condition expressed through a low dynamic range (LDR) image representing the image on the monitor.

Image based relighting. Before the rise of deep learning, attempts were made at non deep learning image based relighting [34,24,33]. Shu *et al.* [34] introduced a face relighting approach that uses a mass-transport formulation for the transfer of illumination between images. Peers *et al.* [24] demonstrated a method for relighting portrait images with flat lighting to match specific target environments, incorporating a reference subject database for approximation. Shih *et al.* [33] adopted a multiscale technique to transfer local image statistics from reference portraits onto new ones, facilitating the matching of attributes like local contrast and overall lighting direction. Recent advancements in deep learning have caused significant shifts to the landscape of portrait relighting [25,11,20]. Research on ratio images for relighting [49,13] was conducted, utilizing public datasets and employing methods based on ratio images. However, this is limited to synthetic data, resulting in a significant domain gap with real data. The widespread application of light stages [5] in gathering data has enabled numerous groundbreaking research endeavors[36,44,38,28]. However, these methods rely on capturing data with a light stage, which are expensive and inaccessible. Instead, our approach builds a personalized relighting algorithm using casually captured videos from the desk recording setup introduced in Sengupta *et al.* [30]. In contrast to Sengupta *et al.* [30], we can collect data during daily computer usage by minimizing numerous constraints, eliminating the need for specific efforts in data collection.

Relighting with explicit decomposition. Creating a virtual relighting dataset through the utilization of synthetic human models to train their networks has been carried out in some studies [31,18,42]. However, when using synthetic data to train a neural network, the large domain gap between synthetic and real data impacts the model's performance on real data. In contrast, others [12,4,27,2,37] generate relit images by using public datasets and employ methods based on 3D model rendering. Specifically, Hou *et al.* [12] takes a more advanced approach by introducing explicit components, where rays originating from the face intersect with other parts of the facial geometry to create relit images. Moreover, the extensive use of light stages [5] has enabled numerous innovative studies [22,40,23,10,43,7,21,47,38,41,19] in this domain. Some researchers have incorporated explicit elements such as albedo, normals, specular maps, and diffuse maps into their methodologies [22,40,23,7]. Others have taken a physics-based rendering approach [10,43] to resolve these issues. Other papers[21,47] aim to manipulate lighting conditions and generate images under different lighting scenarios using texture information. However, explicit decomposition methods require ground truth (GT) data for these intrinsic components to train, either from synthetic data or real light stage data. Obtaining GT data for 'at-home'

captures is impossible, thus we focus on image-based relighting. In contrast, recent studies aim to streamline the capture process, using a mobile phone camera [32] or a sun stage [39] instead of a light stage. Nevertheless, due to their reliance on per-scene optimization, both of these papers lack the capability for real-time relighting and to generalize to unseen appearances of the individual and are only limited to the particular capture. Instead, our approach performs image-based relighting enabling real-time temporally consistent video relighting, at ∼45 fps.

3   Method

Our setup is similar to Gerstner *et al.* [9] and Sengupta *et al.* [30], where a user's face is captured while illuminated by their monitor. By capturing multiple videos of the user's face along with the video on their monitor, we build our 'at home' light stage dataset. We then use these data to train a personalized portrait relighting algorithm that can render the user's face under arbitrary lighting conditions. Specifically, given a portrait image I_src, corresponding source monitor lighting L_src, and target monitor lighting L_trg, our aim is to learn a function G that relights I_src under L_trg:

$$\hat{I}_\mathrm{trg}, \hat{L}_\mathrm{src} = G(I_\mathrm{src}, L_\mathrm{src}, L_\mathrm{trg}; \theta_G). \tag{1}$$

Note that our formulation can be used for scenarios where the source lighting is unknown by simply replacing the input source lighting with the predicted source lighting, unlike previous approaches [30].

   In the following sections, we outline our methodology for portrait video relighting using a monitor as a light stage. Sec. 3.1 outlines strategies for constructing training data pairs from casually captured videos that allow flexibility in facial expression, pose, and ambient lighting. In Sec. 3.2, we introduce our relighting network architecture that disentangles lighting from intrinsic appearance using light-conditioned feature normalization, leading to high-quality relit images. In Sec. 3.3, we propose additional techniques which enforce temporal consistency and eliminate flickering, also using LCFN. Finally, in Sec. 3.4, we discuss how to train our relighting network.

3.1   Constructing training data pairs

While past work [30] imposed requirements of a neutral pose, expression, and dark room, we loosen these constraints to allow subjects in any conditions. The only constraint we maintain is that the room lighting shall not overpower the light emitted from the monitor. For example, if the capture occurs in front of a window with bright sunlight, the light from the monitor will have minimal impact on the face. As in Sengupta *et al.* [30], we aim to generate source and monitor image pairs $(I_\mathrm{src}, L_\mathrm{src})$, as well as target image and target monitor pairs $(I_\mathrm{trg}, L_\mathrm{trg})$, such that we can train our network to produce $\hat{I}_\mathrm{trg}$ where I_trg is the pseudo ground truth. However, due to unrestricted subject movement during data collection, there is a lack of pixel-aligned data, making random pairs
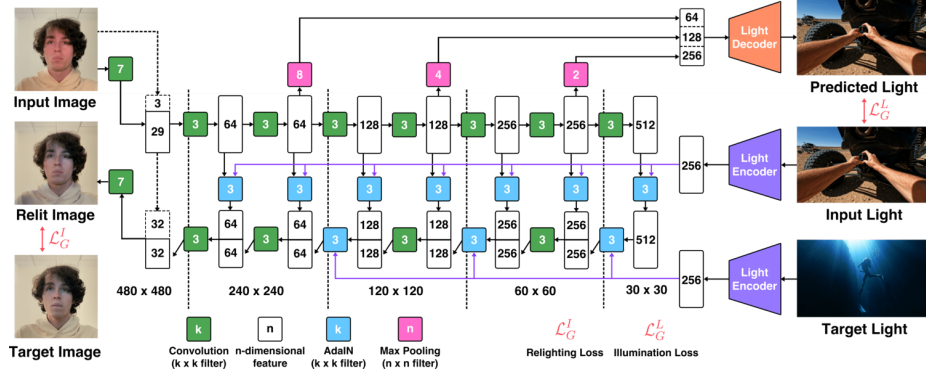
Fig. 3. We first de-light the input image features extracted by the U-Net encoder using Adaptive Instance Normalization (AdaIN) guided by the lighting features extracted from the source lighting with a Light Encoder. We then pass these light-normalized encoder features to the decoder of the U-Net and apply another set of AdaIN guided by the features extracted from the target lighting with the Light Encoder. We additionally predict source lighting from the U-Net encoder using a Light Decoder.

unsuitable. Previous work [30] utilized segmentation for pairing. However, we observed that segmentation is ineffective at finding pairs of images with the same pose and expression. Thus, we instead use facial keypoint detection [16] to obtain source and target image pairs.

3.2   Relighting network architecture

Our network architecture, as illustrated in Fig. 3, is built upon the well-established U-Net [29]. This architecture is comprised of an encoder and a decoder with skip connections, which are commonly used in existing portrait relighting algorithms [36,30,49,13,38]. Our U-Net's encoder, similar to Sengupta *et al*. [30] and Sun *et al*. [36], processes the source portrait I_{src} by applying multiple convolutional layers of varying strides (1 or 2). This process progressively reduces spatial resolution while increasing the number of channels, yielding a latent feature space. The decoder performs the inverse function of the encoder by upsampling from the latent features and simultaneously skip-connecting to intermediate features from the encoder. These skip connections transport high-frequency shape and appearance information from the encoder to the decoder, ultimately resulting in the generation of a realistic relit image. However, they also carry source illumination features from the encoder to the decoder, leading to subpar relighting quality and temporal flickering. To address this issue, we introduce *light-conditioned feature normalization* (LCFN) for the skip-connected features to better disentangle lighting features from intrinsic appearance features.

To disentangle lighting and intrinsic appearance components from the encoded features – i.e. to de-light – we first predict the source lighting from the encoder features. In contrast to Sun *et al*. [36], which predicts the illumination

$\hat{L}_{\text{src}}$ corresponding to the source image I_{src} using the final encoded features, we take a different approach. We extract features at intermediate steps within the encoder, downsample them, and concatenate them using a confidence learning approach [14] to predict the illumination $\hat{L}_{\text{src}}$.

The LCFN module uses the lighting features generated by the lighting encoder to perform Adaptive Instance Normalization (AdaIN) [17] on the encoder features. We begin by using a multi-layer perceptron (MLP) to encode lighting features, transforming the lighting information of L_{src} and L_{trg} into a compact, low-dimensional representation (d = 256). We apply AdaIN to encoder features using the source lighting features, producing normalized features f^l (for $l = 1, \ldots, 7$). Through this normalization process, we induce de-lighting, effectively removing the lighting information present in the encoder features. Starting from the de-lit latent features f^7, we perform progressive bi-linear upsampling. At each upsampling step, we apply AdaIN to the concatenated feature, incorporating the target lighting features encoded by the lighting encoder. This construction using the LCFN module and source lighting prediction allows us to effectively remove source lighting features from the input and only propagate intrinsic appearance features from the encoder to the decoder. We then add target lighting features in the decoder. The LCFN module also contributes towards temporal consistency (see Sec. 3.3). See Fig. 2 for a comparison between our architecture and those of Sun *et al.* [36] and Sengupta *et al.* [30].

3.3   Enforcing temporal consistency

Temporal consistency is vital in making relit videos stable, realistic, and aesthetically pleasing. Previous single-image portrait relighting techniques [30,36] do not incorporate explicit temporal modeling, leading to undesirable flickering artifacts when applied to videos. Accuracy in single-image portrait relighting can often be uncorrelated to temporal flickering. Inconsistencies across frames are even more noticeable when the source lighting L_{src} changes continuously.

When applied to skip-connected features, LCFN provides a natural defense against temporal flickering by removing source lighting features from the input image. However, it cannot ensure temporal consistency on its own. We notice two further problems: (1) when the source lighting gradually changes, LCFN often leaks small amounts of source lighting information to the decoder, leading to flickering; (2) when source lighting changes abruptly, undesirable fading effects can be observed.

To address this issue, we propose a *skip-connected feature smoothing* technique that assumes neighboring frames share the same intrinsic appearance features, obtained after de-lighting input image features with LCFN. We apply a simple exponential smoothing of de-lit features generated by LCFN, denoted as f^l, using all the previous frames:

$$f_t^l := \alpha \cdot f_t^l + (1 - \alpha) \cdot f_{t-1}^l \quad (\text{for } l = 1, \ldots, 7) \tag{2}$$

with $\alpha = 0.7$. Note that exponential smoothing does not work without de-lit LCFN features, which removes time-varying source lighting.

We further notice that when the monitor light changes abruptly the relighting effect is delayed by a few frames, mainly due to the limited refresh rate of the monitor and frame rate of the camera. We thus propose doing a weighted average of source monitor lighting L_{src} from a sequence of previous and current frames to achieve smoother and more natural results:

$$L^t_{\text{src avg}} = \frac{\sum_{i=0}^{N-1} \beta^i L^{t-i}_{\text{src}}}{\sum_{i=0}^{N-1} \beta^i}, \tag{3}$$

where $\beta = 0.6$ and $N = 3$.

3.4   Training relighting network

Our model is trained through minimizing a weighted combination of three loss functions: generator loss, discriminator loss, and monitor loss. The first loss aims to minimize the discrepancies between the true target image I_{trg} in our dataset and the predicted target relit image $\hat{I}_{\text{trg}}$, leading to accurately relit images. We adopted our generator loss (Eq. 4) and our discriminator loss (Eq. 5) from Sengupta *et al.* [30]:

$$\mathcal{L}^I_G = \lambda_{L1}\mathcal{L}_{L1}(I_{\text{trg}}, \hat{I}_{\text{trg}}) + \lambda_P\mathcal{L}_P(I_{\text{trg}}, \hat{I}_{\text{trg}}) \\ + \lambda_C\mathcal{L}_C(I_{\text{src}}, \hat{I}^C_{\text{src}}) + \lambda_D(D(\hat{I}_{\text{trg}}; \theta_D) - 1)^2, \tag{4}$$

$$\mathcal{L}_D = (D(I_{\text{trg}}; \theta_D) - 1)^2 + (D(\hat{I}_{\text{trg}}; \theta_D))^2, \tag{5}$$

where $\mathcal{L}_{L1}$ denotes L1 loss, $\mathcal{L}_P$ denotes perceptual loss [45], $\mathcal{L}_C$ denotes cycle consistency loss [50], and D is the discriminator [15]. $\hat{I}^C_{\text{src}}$ and $\hat{L}^C_{\text{trg}}$ are the outputs from $G(\hat{I}_{\text{trg}}, L_{\text{trg}}, L_{\text{src}}, ; \theta_G)$

The monitor reconstruction loss focuses on minimizing the errors between the predicted source light $\hat{L}_{\text{src}}$ and the true source light L_{src} and is expected to enforce improved disentanglement of lighting information from intrinsic appearance features.

$$\mathcal{L}^M_G = \lambda_{L1}\mathcal{L}_{L1}(L_{\text{src}}, \hat{L}_{\text{src}}) + \lambda_P\mathcal{L}_P(L_{\text{src}}, \hat{L}_{\text{src}}) + \lambda_C\mathcal{L}_C(L_{\text{trg}}, \hat{L}^{trg}_C). \tag{6}$$

Finally, we minimize the image generator loss $\mathcal{L}^I_G$, the discriminator loss $\mathcal{L}_D$, and the illumination loss $\mathcal{L}^M_G$ together:

$$\min_{G,D} \mathcal{L}^I_G + \mathcal{L}_D + \lambda^M_G \mathcal{L}^M_G. \tag{7}$$

4   Experiments

In Sec. 4.1, we first discuss our data collection process, which was based on the approach used to compile the Light Stage at Your Desk (LSYD) dataset. In Sec. 4.2, we perform quantitative and qualitative comparisons with existing single-image portrait relighting algorithms and evaluate their temporal consistency. In Sec. 4.3, we demonstrate predicting, at a low-level, what the user is looking at on the monitor screen. Finally, in Sec. 4.4, we perform ablation studies, evaluating the network architecture's impact on relighting performance.
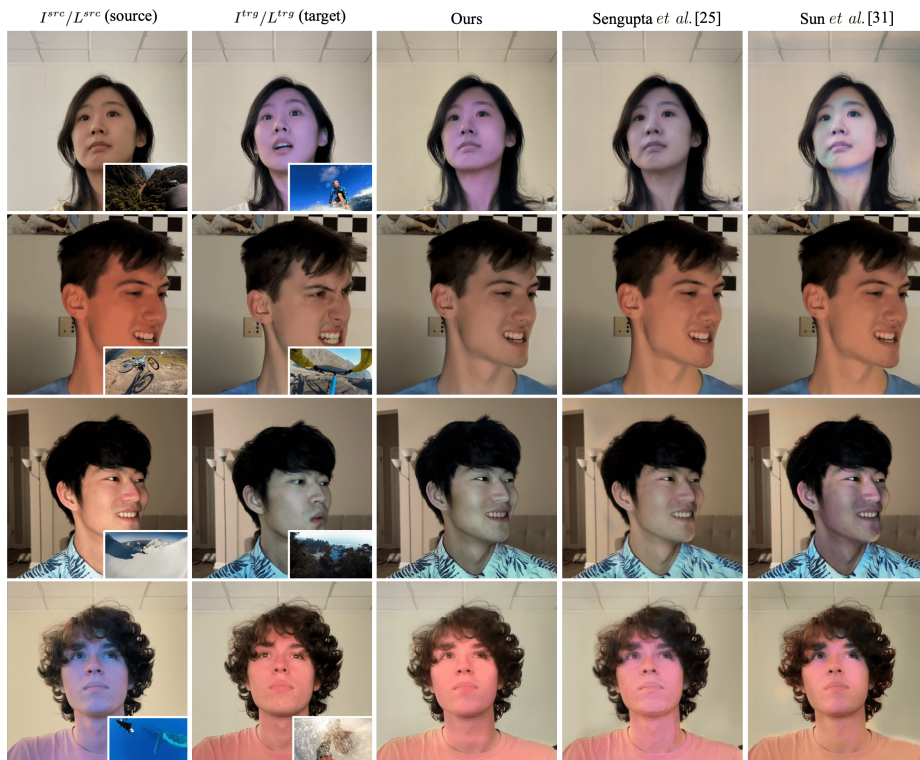
I^{src}/L^{src} (source) I^{trg}/L^{trg} (target)   Ours   Sengupta $et\ al.$[25]   Sun $et\ al.$[31]

Fig. 4. We perform a qualitative comparison with existing techniques [30,36] on the LSYD dataset. Source and target (images & lighting) were unseen during training. All models are personalized, i.e. trained on images of that individual only. We (Col. 3) produce significantly better results compared to existing approaches (Cols 4 and 5).

4.1   Data

We recorded data from 5 users of diverse ethnicities and genders to ensure a wide range of skin types. Each participant wore a variety of outfits, and we used 4 different ambient lighting conditions per person to mimic the conditions of real-life online meetings. We directed the participants to continuously change their facial expression and pose during the capture sessions. Each user's face was captured while watching 8 different videos, each 8 minutes long, on different days with varying appearances. We randomly hold out 1 video for testing and use the remaining 7 videos for training. We use this testing sequence only for qualitative evaluation, not for any quantitative metrics. This is because quantitative evaluation requires a pair of source and target images of the same person in the same pose but under different lighting conditions. This is difficult to obtain accurately for the aforementioned test video sequence since the participants naturally vary their pose and expression over the course of the video. Instead, we capture an additional test sequence, used only for numerical evaluation in

Fig. 5. We perform a qualitative comparison on the OLAT dataset [44]. Our approach outperforms [36,30] and can render strong directional lighting and specular highlights without any explicit modeling of geometry and reflectance.

which the participant is captured in 9 different pose-expression combinations, each with a distinct monitor light. For each pose, we can create $\binom{9}{2} = 36$ source and target pairs as input and pseudo ground-truth, resulting in a total of 324 test data pairs per user. Additionally, we compared methods using Dynamic OLAT Dataset [44] with environment lighting maps [1,8]. During this comparison, we converted HDR environment maps to LDR and utilized a 270° FoV as a monitor light to assess the relighting results. We want to note that all the test data is composed of unseen portraits and lightings for both LSYD and OLAT datasets.

4.2   Comparison with existing approaches

We employ three error metrics to assess relighting performance: RMSE, LPIPS [45], and DISTS [6]. LPIPS and DISTS are more robust to slight differences in pose between the relit image and the pseudo ground truth and detect perceptual differences more effectively than RMSE.

Portrait image relighting. We compared our approach with existing portrait relighting neural architectures — Sun *et al.* [36] and Sengupta *et al.* [30] — by training on our captured LSYD dataset using the same pre-processing for all three architectures (see Sec. 3.1). Our training loss, given in Sec. 3.4, can handle misalignment in source-target pairs in training data, similar to the loss proposed in Sengupta *et al.* [30] (we use an additional loss on source monitor lighting prediction). For Sun *et al.* [36], we train both with their original loss

Table 1. We train a personalized relighting model on 5 users from our LSYD dataset and 4 users from the OLAT dataset [44]. We evaluate these models using source & target portrait images, as well as lighting, that were not encountered during the training phase, and report average RMSE, LPIPS [46], and DISTS[6] scores both on LSYD and OLAT dataset. Our method can perform relighting without source lighting L_{src}, by simply using the predicted light source from our model as input lighting. Our method significantly outperforms Sun *et al.* [36] and Sengupta *et al.* [30].

	Known source lighting	LSYD data			OLAT data [44]		
	$\mathbf{L_{src}}$	LPIPS ↓	DISTS ↓	RMSE ↓	LPIPS ↓	DISTS ↓	RMSE ↓
Sun *et al.* [36] w/ L_{Sun}	–	0.1712	0.1629	8.5958	0.2273	0.1745	6.2692
Sun *et al.* [36] w/ L_{Ours}	–	0.1029	0.1152	8.4476	0.2267	0.1569	6.1898
Ours	–	0.0839	0.0953	8.3222	0.1812	0.1336	6.0931
Sengupta *et al.* [30]	✓	0.1018	0.1105	8.2826	0.2237	0.1675	6.1751
Ours	✓	**0.0832**	**0.0953**	**8.1939**	**0.1809**	**0.1334**	**5.9548**

function L_{Sun} (which expects perfect source-target pose alignment obtained in OLAT data) and with our proposed loss function L_{ours} to specifically handle misalignment in LSYD data. We train personalized models on 5 users from the LSYD dataset and on 4 users from the publicly available Dynamic OLAT Dataset [44] with 2361 indoor environment lighting maps [1,8].

For our quantitative evaluation, we test our model on 1620 test images across 5 users with unseen appearance and lighting conditions on the LSYD dataset and on 7172 test images from the Dynamic OLAT dataset. We present the result in Tab. 1. We observe that our proposed approach outperforms Sengupta *et al.* [30] and Sun *et al.* [36] by 22.3% and 23.6% respectively on the LSYD dataset and by 23.5% and 25.6% on the OLAT dataset, when comparing LPIPS score. Our qualitative comparison, as presented in Fig. 4 and Fig. 5, shows that our model performs superior re-
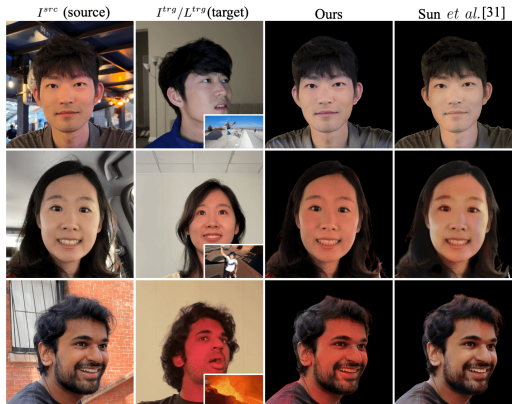

Fig. 6. Our method can also relight portrait images captured with unknown light sources. We relight a source image (Col. 1) with target light shown in the inset in Col. 2. We add a reference of how the face appears under that target light.

lighting in terms of color, quality, and consistency, and can render strong directional lighting and specular highlights without any explicit modeling.

Note that Sun *et al.* [36] does not require the source lighting L_{src} during test time. Our proposed approach can also perform relighting without prior knowledge of source lighting L_{src} by simply predicting $\hat{L}_{src}$ and using it for light-conditioned feature normalization. We show that even in the absence of

L_{src}, our method outperforms Sun *et al.* [36] by 22.6% on the LSYD data and by 25.5% on the OLAT data, in terms of LPIPS.

In Fig. 6, we demonstrate that our approach can relight any portrait image captured 'in-the-wild' without requiring the knowledge of source lighting, and outperforms Sun *et al.* [36]. Our approach first predicts a proxy source monitor lighting, imagining the portrait image to be captured under a monitor lighting, and uses this predicted source lighting in the LCFN module to 'de-lit' the input image and relight it with a target lighting.

In Fig. 7 1st-row, we show the results of relighting an input image with directional lights by moving specific bright areas on the monitor screen. Our method learns to render directional lighting effects and cast shadows as needed. To better illustrate the effect of moving lighting, we first relight the input image with no light reflected from the monitor (row-2 col-1), which produces a relit image under ambient room

Fig. 7. We relight the input image with directional lights (1st row) and without any target light ('Relight A'). We then subtract the 'Relight A' from directional images (row-2 col-2:5). Our method learns to render the effects of directional lighting and decouple ambient and dominant frontal lighting (from monitor).

lighting only. Subtracting this 'de-lit' image from the relit images under directional lighting (row-2 col 2-5) highlights the ability of our method to decouple ambient room illumination from dominant frontal lighting. However, due to the constraint that our lighting is limited to the illumination emitted from the monitor, we may not accurately depict extreme lighting effects. To this end, we show the strong directional light effects using the OLAT dataset in Fig. 5.

Portrait video relighting. Next, we evaluate the temporal consistency of each portrait video relighting algorithm. For each user in the LSYD data, we relit the held-out test video with 50 different target lighting conditions, creating 50 relit videos. We then computed the RMSE between adjacent frames in relit videos as a measure of temporal consistency. Since the pose is almost identical between adjacent frames, lower RMSE error indicates temporally consistent relighting. We then report the average temporal RMSE across all such adjacent frames. In practice,
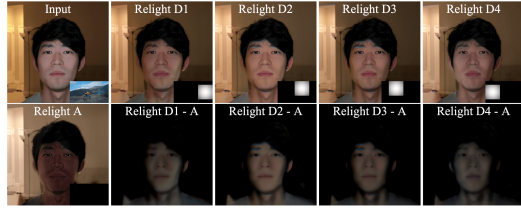
Table 2. We evaluate temporal consistency by relighting a test video captured with varying source lighting with the same target lighting and calculating RMSE between adjacent frames. We then report the average RMSE across all adjacent frames and compute an error rate to indicate the percent of adjacent frames with RMSE higher than a threshold.

	RMSE ↓	Error Rate (%)		
Threshold		>0.2	>0.3	>0.4
Sun *et al.* [36]	5.86	13.53	2.61	1.06
Sengupta *et al.* [30]	6.37	21.83	5.40	1.75
$+L_{\mathrm{src\_avg}}$	5.76	13.31	2.34	0.98
Ours (base)	6.01	16.22	3.61	1.08
$+L_{\mathrm{src\_avg}}$	5.73	13.09	2.31	0.83
+LCFN	5.68	13.04	2.28	0.71
$+L_{\mathrm{src\_avg}}$+LCFN	**5.55**	**12.89**	**2.22**	**0.65**

however, a significant fraction of adjacent frame pairs have extremely similar lighting between the two frames, making their relit frames naturally consistent anyway. Only in a small percentage of adjacent frames does the source lighting significantly change, leading to obvious flickering in the relit video if temporal consistency is not maintained. Thus, in addition to average temporal RMSE, we also compute the error rate for three different thresholds: 0.2 (low), 0.3 (medium), and 0.4 (high), which indicate the percentage of adjacent frames where RMSE error is more than the threshold.

In Tab. 2 and Fig. 8, we compare our approach with and without skip-connected feature smoothing to past works [36,30]. Note that this temporal smoothing of the skip-connected features can only be applied in our framework since we de-light encoder features from the source lighting with LCFN. Both for our approach and for Sengupta et al. [30], we can further apply smoothing of input source lighting to handle abrupt changes. We observe that our method produces the most temporally consistent relighting while also being the most accurate (see Tab. 1). We further note that both skip-connected feature smoothing and smoothing of source lighting improve temporal consistency.
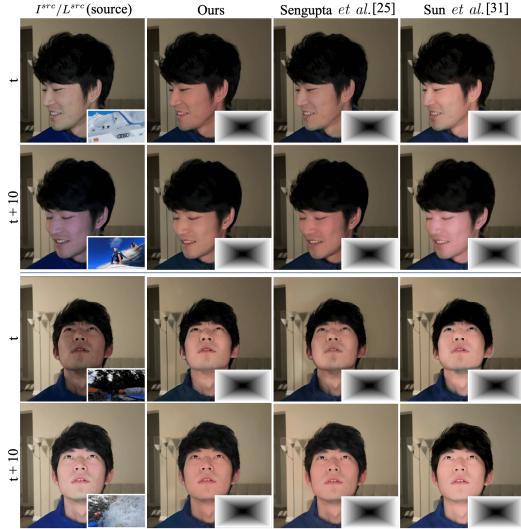

Fig. 8. We show temporal consistency between adjacent frames separated by 0.33s by relighting a test video captured with varying source lighting with the same target lighting. Note that [36,30] both exhibit abrupt changes in lighting between frames t and $t + 10$, while our approach produces a more stable result.

4.3 Monitor prediction

Our network can also be used to predict the source monitor lighting of any input image, as shown in Fig. 9. For the LSYD dataset, we do have images of the source monitor for every input image, so we can calculate monitor prediction accuracy numerically. We observe that our method slightly outperforms [36] with a Mean Absolute Error of 0.15 vs 0.17, also qualitatively
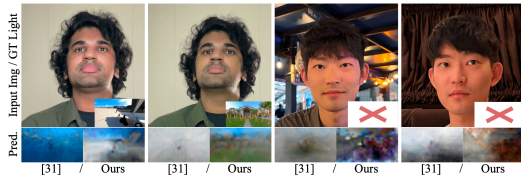

Fig. 9. Monitor prediction comparison on LSYD data (Col. 1 and 2) and "in the wild" setting (Col. 3 and 4). The second row represents the monitor prediction by [36] and Ours.

producing more meaningful visualizations. For portrait images, we do not have ground-truth source lighting, but visualizations show meaningful predictions.

The ability to predict monitor lighting from images captured by our monitor has many implications. This technique can be used to detect deep fake avatars during live video calls by purposefully projecting specific images via screen sharing and observing if we can detect the same image from the webcam feed of other attendees in the call. If there is a mismatch between the projected monitor image and the predicted one, this likely indicates a deep fake avatar in the video call.

4.4   Ablation studies

We report the removal of various components from our relighting network in Tab. 3, specifically LCFN and source monitor lighting prediction using intermediate encoder features. We observe that both improve final relighting performance, which shows their effectiveness in disentangling source lighting information from intrinsic appearance features.

Table 3. Both LCFN and source monitor prediction $\mathbf{L_{src}}$ improve relighting performance by effectively disentangling source lighting information from intrinsic appearance features.

$\mathbf{L_{src}}$	de-lighting	RMSE ↓	LPIPS ↓	DISTS ↓
–	–	8.4230	0.0964	0.1071
✓	–	8.2596	0.0915	0.1013
–	✓	8.1907	0.0904	0.0966
✓	✓	**8.0746**	**0.0853**	**0.0963**

5   Conclusion

We propose a personalized video relighting algorithm that leverages casually captured LSYD data to generate real-time high-quality temporally consistent relit videos under any pose, expression, and lighting conditions. We present a novel network architecture that can perform high-quality relighting on both the LSYD and OLAT datasets, rendering challenging lighting conditions like directional lights, shadows, and specularities. We achieve this without using any 3D information or performing explicit decomposition, simply by achieving better image-based relighting enabled by our proposed neural architecture with LCFN module. Our method enables better lighting quality during live video calls and in portrait images, and produces better harmonization with virtual backgrounds.

Limitation. Since we utilize only the front-facing monitor light as the source lighting, we do not account for scenarios where the light source is located on the sides (90°) or behind (180°). However, it does not necessarily mean that our model cannot learn directional lighting effects, as demonstrated by training on the OLAT dataset which contains full 360° lighting, see Fig. 5.

Ethical considerations. While our primary goal is to allow people to improve their facial appearance with virtual relighting, we note that it is also a form of image manipulation and can be used for malicious purposes [3]. Furthermore, we emphasize that while currently we can only predict the user's direct line-of-sight monitor lighting in low resolution, the potential for high-resolution monitor prediction in the future could raise significant privacy concerns.

References

1. Bolduc, C., Giroux, J., Hébert, M., Demers, C., Lalonde, J.F.: Beyond the pixel: a photometrically calibrated hdr dataset for luminance and color prediction (2023)
2. Chen, Z., Liu, Z.: Relighting4d: Neural relightable human from videos (2022)
3. Choi, J.M., Leung, J., Frahm, N., Christman, M., Bertasius, G., Sengupta, R.: Building secure and engaging video communication by using monitor illumination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4377–4386 (2024)
4. Daichi Tajima, Yoshihiro Kanamori, Y.E.: Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. Computer Graphics Forum (Proc. of Pacific Graphics 2021) **40**(7), 205–216 (2021)
5. Debevec, P., Wenger, A., Tchou, C., Gardner, A., Waese, J., Hawkins, T.: A lighting reproduction approach to live-action compositing. ACM Trans. Graph. **21**(3), 547–556 (jul 2002). https://doi.org/10.1145/566654.566614, https://doi.org/10.1145/566654.566614
6. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. CoRR **abs/2004.07728** (2020), https://arxiv.org/abs/2004.07728
7. Futschik, D., Ritland, K., Vecore, J., Fanello, S., Orts-Escolano, S., Curless, B., Sýkora, D., Pandey, R.: Controllable light diffusion for portraits (2023)
8. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image (2017)
9. Gerstner, C.R., Farid, H.: Detecting real-time deep-fake videos using active illumination. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 53–60 (2022). https://doi.org/10.1109/CVPRW56347.2022.00015
10. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P., Izadi, S.: The relightables: Volumetric performance capture of humans with realistic relighting (2019)
11. He, Y., Xing, Y., Zhang, T., Chen, Q.: Unsupervised portrait shadow removal via generative priors. In: Proceedings of the 29th ACM International Conference on Multimedia. MM '21, ACM (Oct 2021). https://doi.org/10.1145/3474085.3475663, http://dx.doi.org/10.1145/3474085.3475663
12. Hou, A., Sarkis, M., Bi, N., Tong, Y., Liu, X.: Face relighting with geometrically consistent shadows. In: In Proceeding of IEEE Computer Vision and Pattern Recognition. New Orleans, LA (June 2022)
13. Hou, A., Zhang, Z., Sarkis, M., Bi, N., Tong, Y., Liu, X.: Towards high fidelity face relighting with realistic shadows. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
14. Hu, Y., Wang, B., Lin, S.: Fc 4: Fully convolutional color constancy with confidence-weighted pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4085–4094 (2017)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
16. Jin, H., Liao, S., Shao, L.: Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. International Journal of Computer Vision (Sep 2021). https://doi.org/10.1007/s11263-021-01521-4, http://dx.doi.org/10.1007/s11263-021-01521-4

17. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proc. CVPR (2020)
18. Lagunas, M., Sun, X., Yang, J., Villegas, R., Zhang, J., Shu, Z., Masiá, B., Gutierrez, D.: Single-image full-body human relighting. CoRR abs/2107.07259 (2021), https://arxiv.org/abs/2107.07259
19. LeGendre, C., Ma, W.C., Pandey, R., Fanello, S., Rhemann, C., Dourgarian, J., Busch, J., Debevec, P.: Learning illumination from diverse portraits (2020)
20. Liu*, Y., Hou*, A., Huang, X., Ren, L., Liu, X.: Blind removal of facial foreign shadows. In: In Proceedings of British Machine Vision Conference (BMVC). London, UK (November 2022)
21. Meka, A., Pandey, R., Haene, C., Orts-Escolano, S., Barnum, P., Davidson, P., Erickson, D., Zhang, Y., Taylor, J., Bouaziz, S., Legendre, C., Ma, W.C., Overbeck, R., Beeler, T., Debevec, P., Izadi, S., Theobalt, C., Rhemann, C., Fanello, S.: Deep relightable textures - volumetric performance capture with neural rendering. vol. 39 (December 2020). https://doi.org/10.1145/3414685.3417814, http://gvv.mpi-inf.mpg.de/projects/DeepRelightableTextures/
22. Nestmeyer, T., Lalonde, J., Matthews, I., Lehrmann, A.: Learning physics-guided face relighting under directional light. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5123–5132. IEEE Computer Society, Los Alamitos, CA, USA (jun 2020). https://doi.org/10.1109/CVPR42600.2020.00517, https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00517
23. Pandey, R., Orts-Escolano, S., LeGendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., Fanello, S.: Total relighting: Learning to relight portraits for background replacement. vol. 40 (August 2021). https://doi.org/10.1145/3450626.3459872
24. Peers, P., Tamura, N., Matusik, W., Debevec, P.: Post-production facial performance relighting using reflectance transfer. ACM Trans. Graph. 26(3), 52–es (jul 2007). https://doi.org/10.1145/1276377.1276442, https://doi.org/10.1145/1276377.1276442
25. Ponglertnapakorn, P., Tritrong, N., Suwajanakorn, S.: Difareli: Diffusion face relighting (2023), https://arxiv.org/abs/2304.09479
26. Qi, L., Wu, J., Wang, A.N., Wang, S., Sengupta, R.: My3dgen: A scalable personalized 3d generative model (2023)
27. Qiu, H., Chen, Z., Jiang, Y., Zhou, H., Fan, X., Yang, L., Wu, W., Liu, Z.: Relitalk: Relightable talking portrait generation from a single video (2023)
28. Ren, M., Xiong, W., Yoon, J.S., Shu, Z., Zhang, J., Jung, H., Gerig, G., Zhang, H.: Relightful harmonization: Lighting-aware portrait background replacement (2023)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597 (2015), http://arxiv.org/abs/1505.04597
30. 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
31. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6296–6305 (2018)
32. Sevastopolsky, A., Ignatiev, S., Ferrer, G., Burnaev, E., Lempitsky, V.: Relightable 3d head portraits from a smartphone video (2020)
33. Shih, Y., Paris, S., Barnes, C., Freeman, W.T., Durand, F.: Style transfer for headshot portraits. ACM Trans. Graph. 33(4) (jul 2014). https://doi.org/10.1145/2601097.2601137, https://doi.org/10.1145/2601097.2601137

34. Shu, Z., Hadap, S., Shechtman, E., Sunkavalli, K., Paris, S., Samaras, D.: Portrait lighting transfer using a mass transport approach. ACM Trans. Graph. **37**(1) (oct 2017). `https://doi.org/10.1145/3095816`, `https://doi.org/10.1145/3095816`
35. Song, G., Cham, T.J., Cai, J., Zheng, J.: Half-body portrait relighting with over-complete lighting representation (06 2021)
36. 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`
37. Sun, T., Lin, K.E., Bi, S., Xu, Z., Ramamoorthi, R.: Nelf: Neural light-transport field for portrait view synthesis and relighting (2021)
38. Sun, T., Xu, Z., Zhang, X., Fanello, S.R., Rhemann, C., Debevec, P.E., Tsai, Y., Barron, J.T., Ramamoorthi, R.: Light stage super-resolution: Continuous high-frequency relighting. CoRR **abs/2010.08888** (2020), `https://arxiv.org/abs/2010.08888`
39. Wang, Y., Holynski, A., Zhang, X., Zhang, X.: Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20792–20802 (2023)
40. Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F.: Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Trans. Graph. **39**(6) (Nov 2020). `https://doi.org/10.1145/3414685.3417824`
41. Yang, H., Zheng, M., Feng, W., Huang, H., Lai, Y.K., Wan, P., Wang, Z., Ma, C.: Towards practical capture of high-fidelity relightable avatars (2023)
42. Yeh, Y.Y., Nagano, K., Khamis, S., Kautz, J., Liu, M.Y., Wang, T.C.: **41**(6), 1–21 (nov 2022). `https://doi.org/10.1145/3550454.3555442`, `https://doi.org/10.1145`
43. Yeh, Y.Y., Nagano, K., Khamis, S., Kautz, J., Liu, M.Y., Wang, T.C.: Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. ACM Transactions on Graphics (TOG) (2022)
44. Zhang, L., Zhang, Q., Wu, M., Yu, J., Xu, L.: Neural video portrait relighting in real-time via consistency modeling (2021)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595. IEEE Computer Society, Los Alamitos, CA, USA (jun 2018). `https://doi.org/10.1109/CVPR.2018.00068`, `https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068`
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CoRR **abs/1801.03924** (2018), `http://arxiv.org/abs/1801.03924`
47. Zhang, X., Fanello, S.R., Tsai, Y., Sun, T., Xue, T., Pandey, R., Orts-Escolano, S., Davidson, P.L., Rhemann, C., Debevec, P.E., Barron, J.T., Ramamoorthi, R., Freeman, W.T.: Neural light transport for relighting and view synthesis. CoRR **abs/2008.03806** (2020), `https://arxiv.org/abs/2008.03806`
48. Zhang, X., Barron, J.T., Tsai, Y.T., Pandey, R., Zhang, X., Ng, R., Jacobs, D.E.: Portrait shadow manipulation. vol. 39 (2020)
49. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.: Deep single-image portrait relighting. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7193–7201 (2019). `https://doi.org/10.1109/ICCV.2019.00729`
50. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)