Temporally Consistent Relighting for Portrait Videos

Sreenithy Chandran[1], Yannick Hold-Geoffroy[2], Kalyan Sunkavalli[2], Zhixin Shu[2], and Suren Jayasuriya[1]

[1]Arizona State University
[2]Adobe Research

Abstract

Ensuring ideal lighting when recording videos of people can be a daunting task requiring a controlled environment and expensive equipment. Methods were recently proposed to perform portrait relighting for still images, enabling after-the-fact lighting enhancement. However, naively applying these methods on each frame independently yields videos plagued with flickering artifacts. In this work, we propose the first method to perform temporally consistent video portrait relighting. To achieve this, our method optimizes end-to-end both desired lighting and temporal consistency jointly. We do not require ground truth lighting annotations during training, allowing us to take advantage of the large corpus of portrait videos already available on the internet. We demonstrate that our method outperforms previous work in balancing accurate relighting and temporal consistency on a number of real-world portrait videos.

1. Introduction

Portrait videos are a large portion of user content being uploaded daily to online and social media platforms. They typically feature one person—such as a news anchor, an entertainer, or a communicator—whose face and upper torso are featured prominently. While relatively simple in concept, high-quality portrait videos are hard to capture, requiring proper illumination equipment and a controlled environment to get the right aesthetics.

The rise of user-created amateur content and video conferencing has brought its share of portrait videos shot with relatively modest cameras and lighting equipment (e.g., ring lights, cellphone cameras). Illumination can be challenging to control fully in these situations, and manually post-processing each video to correct its lighting requires training and is time-consuming. Recent advances allow for post-capture editing of lighting in single portrait images [33, 48], but produce flickering results when applied to videos on a frame-by-frame basis. Extending these fully supervised methods to work directly on videos instead of images proves to be prohibitively expensive and laborious, as they require portraits with a vast diversity of illumination conditions for training. This data is not available publicly and is usually acquired using expensive lighting and capture rigs such as light stages [13, 26] or mechanical gantries [8, 12].

In this work, we propose an end-to-end differentiable portrait video relighting pipeline that generates temporally consistent videos. Our pipeline consists of a portrait relighting method [33, 48] and a video consistency method [2, 20, 18] trained jointly using three losses: temporal loss, perceptual loss, and lighting loss. In addition to the pipeline, we also introduce a novel method to train it involving the generation of relit portraits on the fly, thereby eliminating the need for lighting annotations during training. Thus, our method can be trained using any existing video dataset.

We summarize our contributions as the following:

- An end-to-end differentiable pipeline that involves the following: facial alignment and segmentation, colorspace conversion, single image portrait relighting, blending of the face back onto the upper body, blind consistency with temporal, lighting and perceptual consistency losses.
- A novel method for training this pipeline with readily available portrait videos, without ground truth lighting annotation needed.
- Experiments on a variety of portrait videos demonstrating state-of-the-art relighting accuracy while preserving temporal consistency better than existing non-lighting-aware methods, as demonstrated by our user study.

Despite the promising results we obtain, there are future directions that could enhance our method. First, our pipeline only handles human faces and does not relight the background. Full scene relighting from a single image is not yet tackled in the literature. Furthermore, some blending artifacts are slightly visible, which could be alleviated by

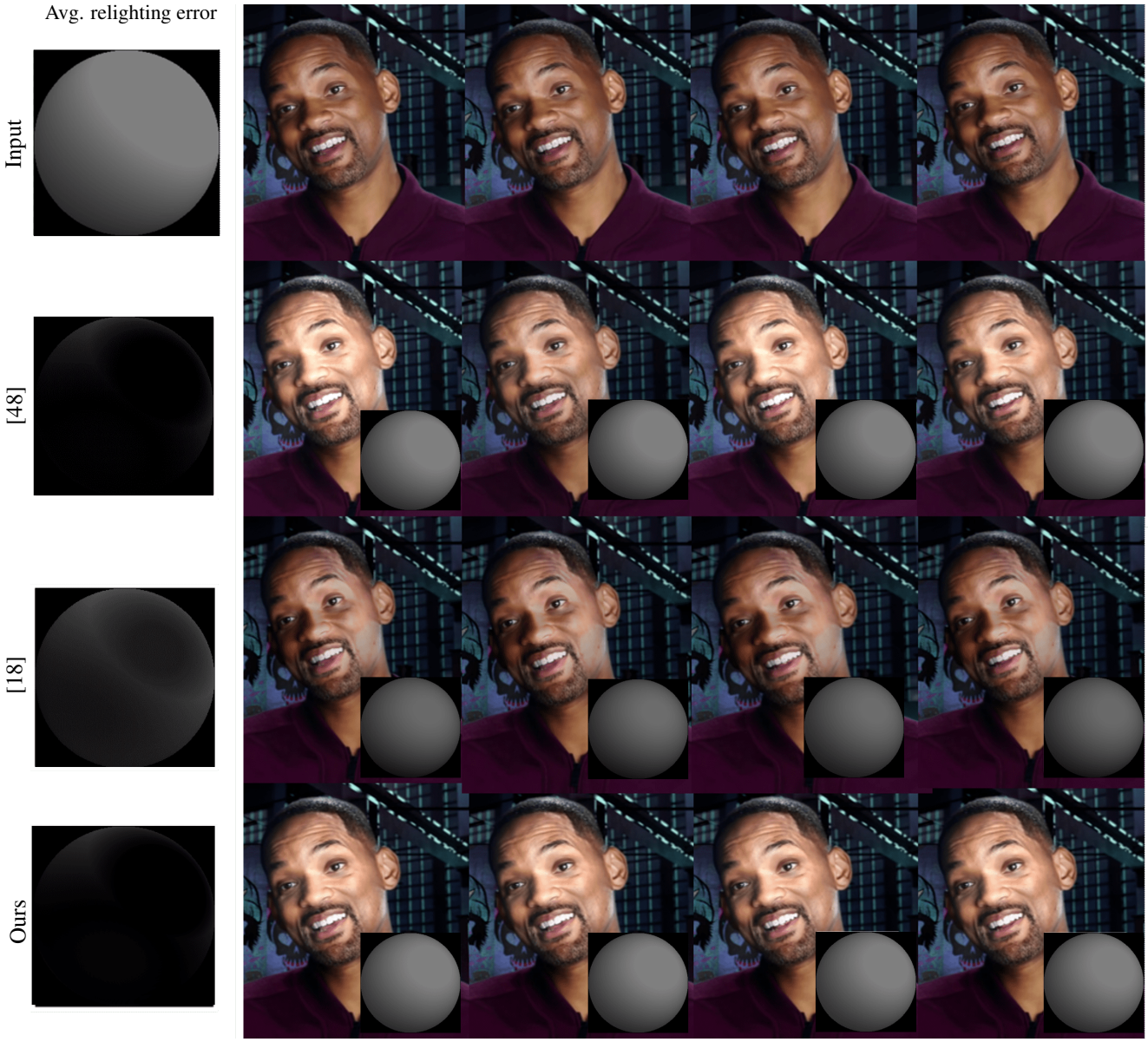

Figure 1. Given input video frames (top), we apply a single image relighting methods per-frame [48] (2nd row), producing a video that has the desired target lighting on average—as shown by the low average lighting estimation error, visualized via rendered spheres in the leftmost column—but flickers between frames. Applying a blind video consistency method [18] (3rd row) removes the flickering, but also significantly changes the lighting. In contrast, our method (bottom) produces a flicker-free video that preserves the desired target lighting. Note, the first column is the average relighting error, the darker the image, the better the corresponding method.

leveraging recent compositing methods such as [44].

2. Related Work

While the problem of end-to-end portrait video relighting has not been tackled in the past, there has been significant work on related topics. In the following, we discuss relevant steps such as human face modeling, lighting estimation, relighting, and video temporal consistency.

Face modeling and editing in the wild: The problem of face detection and alignment has been under scrutiny for many years, first using optimization techniques [14, 5] and more recently with deep learning-based methods [49, 45, 11, 9]. Once detected, inverse rendering methods can be applied to model a face, a topic initiated by the 3D Morphable Model [1] which demonstrated successful face relighting under specific circumstances. Shu et al. [31], followed by [34], extend the 3DMM concept using a generative adversarial network to decompose portraits into reflectance prop-

erties, including shape, albedo, and lighting, from which they perform lighting transfer.

As most face modeling methods present in the literature are tailored for single images, they can be naively extended to video by processing each frame independently. Methods to automatically select the most representative frame for the face were later developed [23]. Very recently, work has been proposed to learn to directly act on the face mesh, providing expression synthesis for videos [28]. This concept of modeling expressions has also been used in the context of video compression [36], considerably reducing the bandwidth required during videoconferencing sessions.

Lighting estimation and relighting: A controlled lighting environment such as a light stage [8, 7] allows for the acquisition of face reflectance properties with high fidelity. Once the properties are captured, it is possible to use them to relight a facial performance in post-production [27]. Similarly, Einarsson et al. [10] develop a technique that uses time-multiplexed lighting coupled with high-speed cameras to capture a person running on a treadmill. However, these techniques require complex lighting setups, which can be expensive or inconvenient to use. To circumvent this, portrait lighting estimation methods [19, 4] propose to use the human face as a light probe, estimating the environment lighting from a single image. Furthermore, lighting estimation was also proposed for videos of generic scenes [22], taking advantage of the entire sequence of frames.

In addition to lighting estimation, relighting has also received much attention from the computer vision community. Techniques using inverse rendering methods on a coarse geometry of the face were initially proposed [39, 37]. Despite the promising results offered by those techniques, their use of a face geometry proxy limits the relighting to a fixed region of the portrait, precluding its use on hair. To mitigate this, image-based methods using additional hardware such as an IR projector [35] were proposed. [40] performs relighting of videos but requires an input video with uniform illumination across the face. Additionally, their lighting transfer scheme requires a similar skin color and face geometry between the target and reference videos, limiting its applicability in the wild. Recently, deep learning-based methods propose in-the-wild relighting for humans bodies [16], or even entire scenes [50, 21, 43], allowing a user to change the lighting of the whole image post-capture. Deep learning methods have also been applied to single image portrait lighting manipulation. Zhang et al. [47] automatically remove cast shadows and simulate a fill light to dampen stark lighting and improve the visual appeal of portraits. Single image portrait relighting methods [38, 48, 33, 24] change the lighting to a user-specified target lighting condition. These methods are all trained with datasets (either real or synthetic) of photographs with ground truth lighting annotations. We extend such methods—in particular, Deep Portrait Relighting [48]—to videos *without* requiring such annotation video relighting data.

Video temporal consistency: A naive way to extend single-image methods to video is to apply them on a frame-by-frame basis. Doing so typically results in videos with temporal discontinuities and noticeable flickering. This problem was originally addressed by Blind Video Temporal Consistency [2] using a gradient-domain technique. Since then, multiple methods using a CNN [18], a GAN [6], leveraging the Deep Video Prior [20], or specifically tailored for full-body human synthesis [41] were proposed. In our work, we draw inspiration from Lai et al. [18] to develop a fast feed-forward network that enforces the video's temporal consistency. However, when directly applying [18] on relit videos, we observe a drift in the lighting, straying away from the target lighting specified. To solve this, we explicitly integrate lighting cues to the temporal consistency model, allowing for a flicker-free video that preserves the target lighting defined by the user throughout its duration.

3. Approach

A straightforward way to extend single image relighting methods to video is to apply them on each frame independently. In this work, we leverage the Deep Portrait Relighting (DPR) network from [48] to perform re-illumination. We want to point out that our method is not tied to this specific method and can extend any image relighting method. We encourage the reader to look at our supplementary video results to better appreciate the results obtained by applying DPR on a per-frame basis. In general, doing so generates high frequency flickering and causes global changes in the lighting level and average color of the face. These problems are the motivation for our proposed system to improve the temporal consistency of the lighting in portrait videos.

Problem Challenges and Assumptions: There are several key challenges to overcome with designing an end-to-end portrait video relighting system. Our system needs to generate temporally consistent lighting across frames, but has to be robust to significant variations in face geometry, reflectance, and pose. In particular, a moving face in the video requires accurate facial alignment to ensure the resulting relighting tracks the movement and is robust to occlusion. Further, the relit face must be seamlessly blended back onto the upper torso, neck, and hair without artifacts.

We make some critical assumptions in our pipeline to make the problem of video relighting more tractable. In this paper, we focus facial relighting and do not solve the harder problem of relighting in general. Our method will relight regions close to the facial region such as the upper neck and hair next to the forehead, but will not perform full human body relighting nor change lighting on long hair. Despite this, we believe modeling and editing

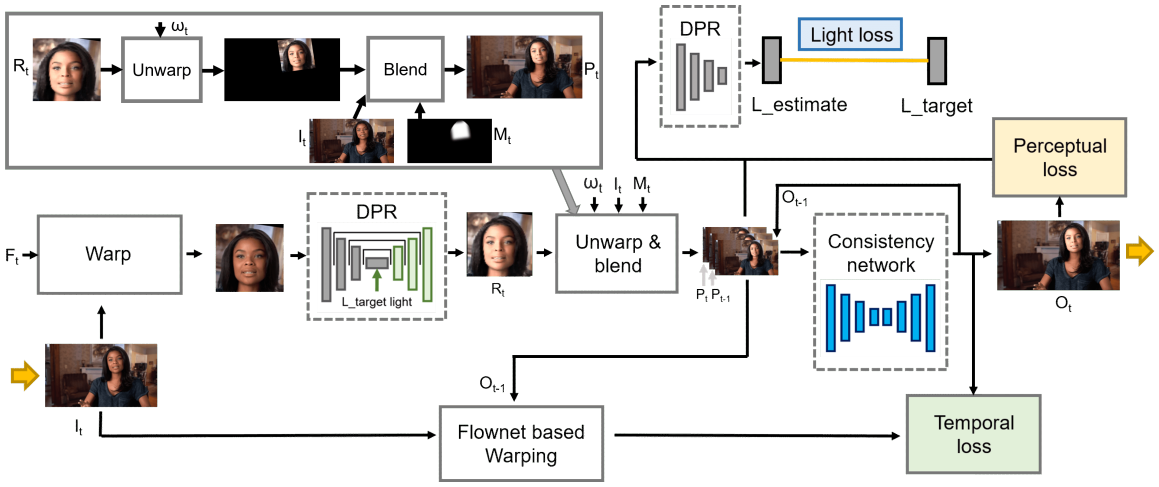

Figure 2. **Pipeline of the proposed Method.** Our end-to-end relighting framework consists of two major sections, the relighting & the consistency part. The input video frame I_t is relit by the portrait relighting architecture, these processed frames P_t is then passed through the consistency network along with the previously stabilised relit frame O_{t-1} to obtain the temporally consistent relit image O_t. For a video segment, the first relit image is taken as O_{t-1} for $t = 0$. The network is trained with loss functions to account for temporal consistency, data fidelity and lighting preservation. For details regarding the architecture, please refer to the supplementary.

face lighting on videos is a useful first step. In addition, the quality of our relit results is based on Deep Portrait Relighting (DPR), which uses spherical harmonics to relight faces. As such, our method inherits its limitation of capturing only low-frequency lighting. However, our generic pipeline can be applied to a different relighting component and exploit a more powerful lighting representation with minimum changes. Finally, we assume the target lighting is static during the entire video, so that we can properly constrain the temporal consistency of this lighting. Dynamic lighting with a dynamic moving face is a much more challenging problem that is out of scope of the current work.

4. Proposed System

Our proposed pipeline, schematized in Figure 2, introduces a joint network architecture coupled with strong loss constraints specifically for video relighting. In the following, we first describe how we perform face alignment, followed by our relighting step. Then, we detail our compositing step, and finally the consistency network with which we balance temporal and lighting consistency losses.

4.1. Facial Alignment and Warping

To execute our method, we employ an input video frame I_t which we relight using DPR [48] and denote R_t. We employ the same portrait pre-processing steps as DPR, as described in [17]. Concretely, we use the method of Bulat et al. [3] to detect facial landmarks and warp the face to some specific location in the image. In practice, we cache the warped images and their masks to reduce computational overhead during training.

4.2. Portrait Relighting and Blending

To relight the face in the input frame, we make use of Deep Portrait Relighting (DPR) [48]. This network is inspired by the Hourglass architecture [25], which is composed of an encoder and a decoder connected together by a bottleneck layer. DPR changes slightly this architecture by adding a subnetwork connected at the bottleneck layer. The role of this subnetwork is twofold: it estimates spherical harmonic coefficients from the encoder—performing lighting estimation from the input image—and encodes the input target lighting for the decoder. In summary, this network takes as input a source image, a target lighting and outputs the image R_t relit under the given lighting condition. Note that the relighting is done on the luminance channel of the LAB image. For specific details regarding implementation, we encourage the reader to refer to the original paper [48].

The relit image R_t is then un-warped to the original video frame using the inverse of the warping operation described in §4.1. In addition, we compute a background mask M_t using [42], which we feather lightly using a Gaussian blur operation. We then blend the relit image R_t onto the original video frame I_t using the mask M_t using

$$P_t = M_t \odot R_t + (1 - M_t) \odot I_t \ , \tag{1}$$

where $\odot$ is element-wise multiplication. To make this pipeline end-to-end, all the processing such as warping, color space conversion, and blending are implemented as differentiable functions using Kornia [29]. At this point, we have pairs of images (I_t, P_t) that represent the same video frame lit under two different illumination conditions.

4.3. Consistency Network

Blind consistency methods in video typically utilize self-supervision to mitigate temporal artifacts for processed video. In our design exploration, we first applied a blind consistency network [18] directly to the output of a DPR-processed video. This consistency network comprises of convolutional layers followed by residual blocks. The output of the residual block is then passed through a ConvLSTM [30] layer. This ConvLSTM layer learns relationship between neighbouring frames to give a temporally consistent output. The current output frame is fed back as an input to the consistency network for next pair of frames being processed.

We incorporate this same consistency network architecture into our pipeline. As shown in the Fig. 2, we obtain the stabilized output frame O_t from the consistency network Ψ_θ with trained parameters θ with

$$O_t = \Psi_\theta \left(P_t, P_{t-1}, O_{t-1} \right), \tag{2}$$

where P_t and P_{t-1} are the relit frames at time t and $t-1$ respectively (see §4.2), and O_{t-1} is the previously stabilized relit frame at time $t-1$ output at a previous iteration. For the first image to be stabilized $t = 2$, we use $O_{t-1} = P_t$.

Applying [18] successfully mitigates the temporal inconsistencies but detrimentally changes the content of the video as time progresses. For instance, the lighting seems to be decoupled from the facial geometry and shadows and bright reflectance patches across the face seem to not conform to the facial movement in the video. Further, there is a subtle color warping or drift of both the face and background over time. To circumvent those issues, we introduce the additional loss functions described in the following.

4.4. Loss Functions

The primary goal of the network pipeline is to perform relighting in a consistent manner, while preserving data-fidelity and ensuring that relit videos maintain the same requested target lighting. For this purpose we use the following loss functions.

Lighting Estimation Loss: The first loss function's primary goal is to ensure that the consistency network does not alter the lighting as time progresses during the video, a phenomena we observe when applying blind consistency to the output of DPR-relit video (see Fig. 3).

To prevent this, we utilize the light estimation network from the DPR network to perform a lighting estimation loss. In other words, we require the lighting estimated from each video frame to be temporally stable and not shift around. To implement this, the output frame O_t is passed through the encoder of the relighting network to extract the SH of the light associated with it, let us denote this as $\mathcal{L}_{\text{light}}$. The error is then calculated on the 9 SH coefficients L^i as

$$\mathcal{L}_{\text{light}} = \sum_{i=1}^{9} \left(L^i_{\text{target}} - L^i_{\text{estim}} \right)^2 . \tag{3}$$

Temporal Loss: The second loss enforces the temporal consistency on the processed video. We follow [18] and implement this loss on both short and long-term. The short-term loss is computed by measuring the warping error between two consecutive frames, while the long-term consistency is applying the warping error between the first frame of the video and the current frame. The optical flow component of the warping error is provided by FlowNet2 [15] being processed on the aligned face images. Formally, the temporal loss is defined as

$$\mathcal{L}_{\text{temporal}} = \sum_{t=2}^{T} \sum_{j=1}^{N} M^j_t \left(\left\| O^j_t - \hat{O}^j_{t-1} \right\|_1 + \left\| O^j_t - \hat{O}^j_1 \right\|_1 \right), \tag{4}$$

where O^j_t is the j^{th} pixel of the t^{th} output of the consistency network, $\hat{O}_{t-1}$ is the frame O_{t-1} which was warped using the optical flow estimation, and M^j_{t-1} is the optical flow uncertainty mask given by $M_t = e^{-50 \left\| I_t - \hat{I}_{t-1} \right\|_2^2}$. This uncertainty mask prevents penalizing the network for errors in the optical flow estimation.

Perceptual Loss: The final loss we implement is a perceptual loss based on VGG features [32], computed as

$$\mathcal{L}_{\text{perceptual}} = \sum_{t=2}^{T} \sum_{j=1}^{N} \left\| \phi_4(O^j_t) - \phi_4(P^j_t) \right\|_1 , \tag{5}$$

where $\phi_4(\cdot)$ is the output of the 4th layer (i.e., **relu4-3**) of the pretrained VGG-19 network. This allows the network to ensure the content of the video is preserved including texture, facial movement and occlusions that occur.

We train our model by summing the three aforementioned losses as:

$$\mathcal{L} = \lambda_l \mathcal{L}_{\text{light}} + \lambda_t \mathcal{L}_{\text{temporal}} + \lambda_p \mathcal{L}_{\text{perceptual}} , \tag{6}$$

with various weights $\lambda_\cdot$ to balance the losses during training. In our experiments, we empirically use $\lambda_l = 5$, $\lambda_t = 100$ and $\lambda_p = 10$.

5. Implementation and Evaluation

Dataset: Our dataset consists of 190 videos ($\approx$ 90000 frames) of interviews and recordings of celebrities, which we split into 80% training and $20\%$for our validation. These videos of varying lengths feature significant variations in skin tone, facial geometry, pose/movement, facial expressions, and background environments. Additionally, we prune the videos containing multiple subjects before carrying out the preprocessing. For each video, we extract facial

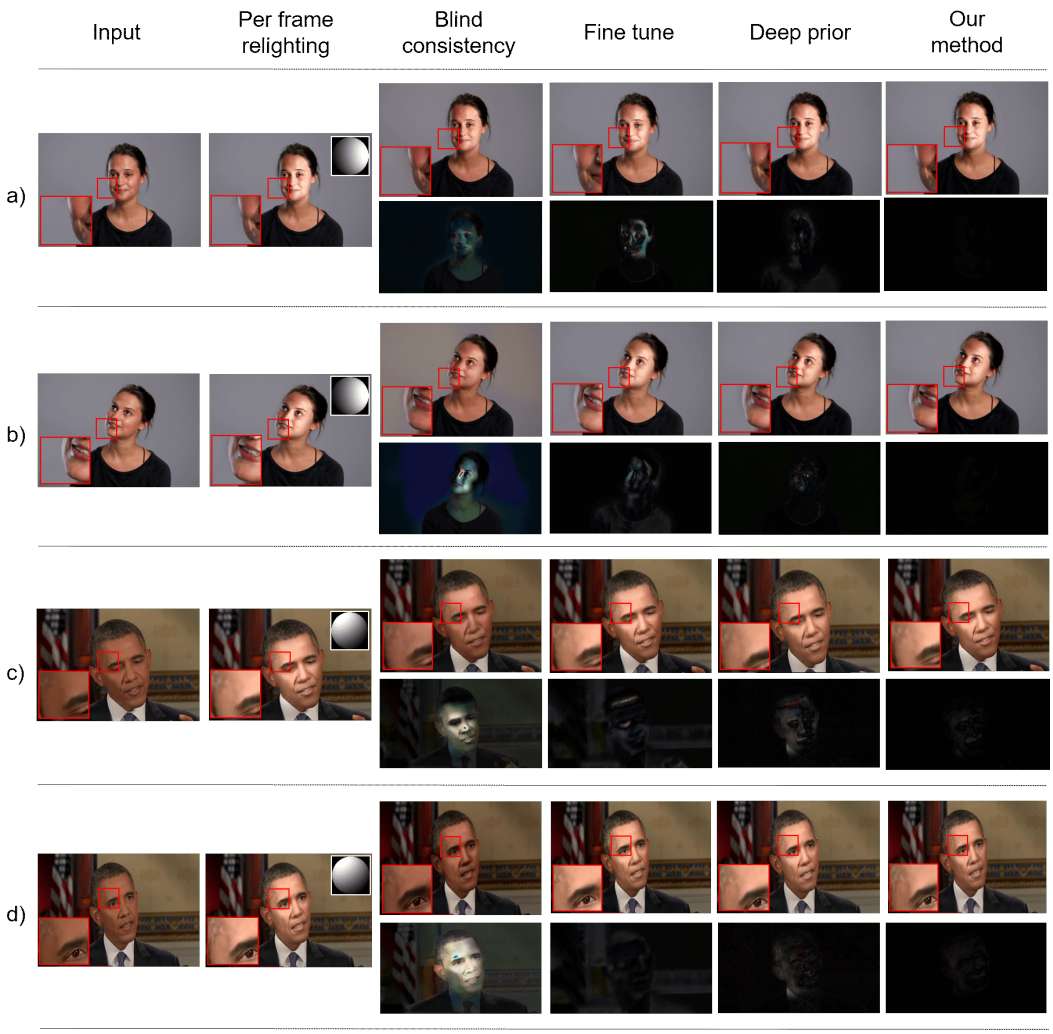

Figure 3. Qualitative comparison on relighting, showing the error (below) when comparing the relit frames (above) to the target lighting. This evaluates the accuracy of each method at preserving the image content. The difference images (rows 2, 4, 6, 8) are multiplied by 3.

landmarks F_t, warping matrix ω_t, and background masks M_t during preprocessing. For details regarding the training procedure, please refer to the supplementary material.

Evaluation Metrics: To evaluate our methods quantitatively, we employ traditional temporal metrics commonly used for video consistency, including the *Warping Error* from [18]. In addition, we evaluate the data fidelity in terms of *Perceptual Similarity* metric using LPIPS [46].

Comparisons: Since, to the best of our knowledge, no end-to-end portrait video relighting method exists, we develop our own baselines for comparison. In addition to utilizing DPR per frame, we also compare against the pipeline of DPR + blind consistency[18]. In addition, we also fine-tune blind consistency on our training data to help demonstrate the need for our additional loss functions. Another related but slightly different approach in the literature is the use of deep priors [20] to estimate lighting per frame. We compare explicitly against the method in [20] to show the differences between a deep prior and a trained end-to-end architecture.

Training details: Our algorithm was trained on two Geforce GTX 1080Ti GPUs. We use ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of $1e-04$. During training, the sequence length is fixed to 11, with the first frame chosen as the reference. We trained the model for 130 epochs, which took over 50 hours.

6. Experimental Results

In this section, we present our experimental results in both qualitative and quantitative comparisons. **We highly encourage the reader to view the supplementary video to grasp our method's capabilities better.**

In Figure 3, we show two frames from two different portrait videos, followed by the results from DPR [48], pre-

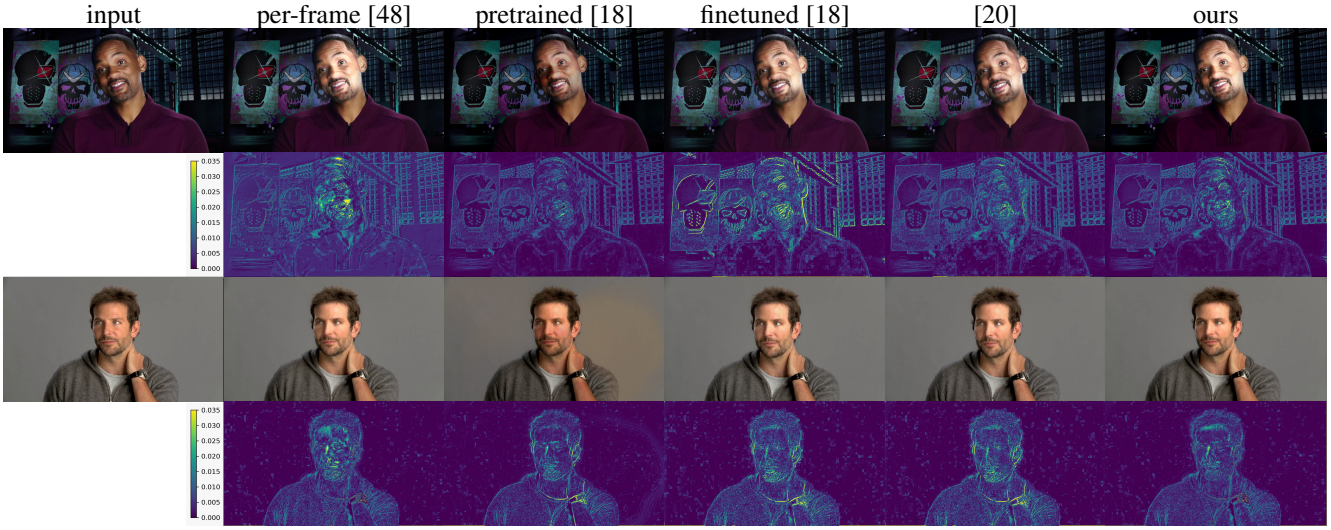

Figure 4. Qualitative warping error evaluation. We relight a video frame (leftmost) using different methods (rows 1,3), and show the warping error between two consecutive frames (rows 2,4, see §4.4 for more details). Note how DPR applied per-frame can generate images with a lot of flicker, as seen in the warping error (row 2,4 left). [18] generally removes flicker very well (low warping error), but loses the lighting content and introduces a brown-ish tinge (see Fig. 3). [18] and [20] perform generally well but tend to lose high-frequency details. Our method (rightmost) provides a good balance between temporal consistency and content preservation.

trained blind consistency [18] applied to the DPR relit video with and without finetuning, deep prior method [20], and finally our method. Underneath each method is the difference image between that frame and the reference relighting produced by DPR. This helps visualize any errors in lighting in the frame, effectively measuring the method's capability to keep the lighting content stable. As we can see, the blind consistency methods (both pretrained and finetuned) have significant errors in their lighting as compared to the reference relit frame. This is due to the method's focus on minimizing temporal flickering artifacts over keeping the physically-plausible lighting and colors. We also noticed that the pretrained blind consistency model would experience a drift in colors slowly throughout the video, as shown in the second frame relative to the first in Figure 3.

We further investigate how well each method preserves lighting by using DPR to estimate the lighting in each frame of the relit video. As shown in Figure 5, the blind method's lighting is changing distinctly over time as compared to the reference relit target. Both the deep prior and our method produce frames with more accurate lighting.

In Figure 4, we show the average warping error on two other portrait videos. This warping error is computed in two steps. First, we compute the optical flow between frames of the original video. Second, we warp the adjacent frames of each relit video using this optical flow and then compute the difference between them. Note how the blind consistency methods achieve the lowest warping error. This measure corroborates the viewing experience and adequately captures our perception of the amount of temporal flickering in the video, as the blind consistency methods featured the least amount of flickering of all videos. However, the results produced by this technique progressively change color and lighting throughout the video. Both Figures 3 and 4 demonstrate that our method achieves a good compromise in preserving the lighting while minimizing a large amount of the temporal flicker.

To quantify the previous observations, we present the results of our error metrics for the various methods in Table 1. These numbers are evaluated on 5 randomly chosen videos, relit with 5 different lighting conditions. Our method achieves the best average LPIPS distance [46] of all methods with 0.0028, while proposing a competitive average warping error of 0.0089, much better than the error of 0.012 from the frame-by-frame relit video. As mentioned before, both blind consistency methods change their content over time, which results in a higher LPIPS error. We note that our method is on-par with these methods with respect to warping error, lending evidence to our method as satisfying both constraints for portrait video relighting well.

6.1. User Study

To further emphasize the importance of temporal stability for humans, we conducted a user study to investigate how the subjects perceive our method against per-frame DPR, Blind Consistency, Finetune and Deep Prior. Twenty-five subjects participated in the study (12 females, 13 males). None of the participants were aware of the research or the methods. All the participants had a normal or corrected-to-normal vision. We utilized 10 videos relit by

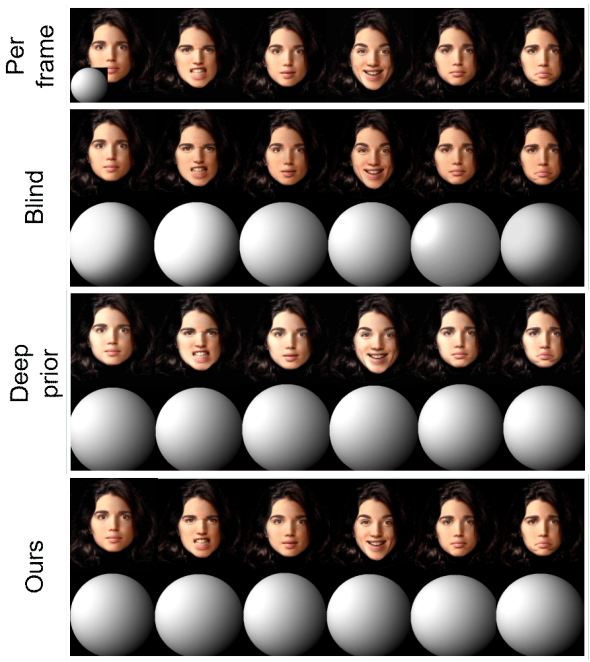

Figure 5. Starting from the top, the rows correspond to [48],[18],[20] and our method consecutively. Each row shows the frames of a video as handled by one of the aforementioned methods. The inset in the per-frame row, shows the target lighting used for the relighting. Applying blind temporal consistency (row 2) to relit portrait video significantly alters the lighting condition from frame to frame (shown via spheres rendered with the lighting). Our approach (row 4) ensures a consistent lighting condition throughout the frames of the video.

Method	Warp Error ↓	LPIPS [46] ↓
Per-frame DPR [48]	0.012160	-
Blind Consistency [18]	**0.007050**	0.01496
Finetune [18]	0.008817	0.00635
Deep prior [20]	0.009143	0.00514
Our Method	0.008922	**0.00276**

Table 1. Quantitative evaluation comparing state-of-the-art methods on flickering (warping error between adjacent frames) and lighting and content preservation (LPIPS between the per-frame DPR result), both lower is better. Our method preserves the best the lighting while providing competitive flicker removal, much better than naive per-frame relighting.

all the methods, and we limited all videos to 30s sequences.

Study Details: The study was a *two-alternative forced-choice*. Given two differently relit videos, the subject was forced to select one choice. Due to COVID-19 restrictions, the users were asked to perform the study on their respective mobile devices. Each trial compared our result to either (per-frame DPR, Blind Consistency, Finetune, and Deep Prior), where our result is randomly set as either the first or second video. Each video is displayed for 30 with a 2s blank screen between each. The users were instructed to base their decision on what they found most appealing aesthetically and was not straining their eyes. Each user had to undergo 20 trials, that is 5 trials per pair of the 4 combinations. The users could pause the study and resume at their convenience.

Results: The user study results are shown in Fig. 6. Despite slight artifacts present in our results, our method is overwhelmingly preferred over all other methods, including in every case over DPR and 74% of the time over finetuned blind consistency, which is the next best performing method after our method.

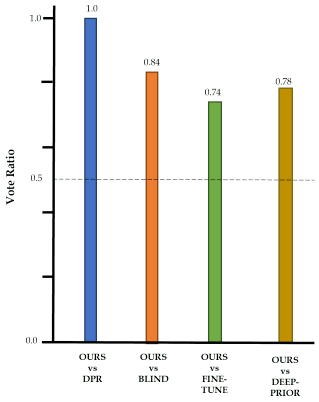

Figure 6. User study results showing the vote ratio comparing two methods at a time. 1.0 denotes humans preferred the first method throughout all trials. The very high vote ratio ($\gg 0.5$) shows a clear preference for our method.

7. Discussion

In this paper, we tackle the challenge of extending a deep single portrait image relighting method to video. We introduce a novel pipeline that synergizes single-image portrait relighting with blind video consistency. To the best of our knowledge, this is the first end-to-end video relighting pipeline proposed for portrait videos. We hope our work sparks interest in portrait video relighting and leads to the production of high-quality videos without the need for time-consuming processes or expensive equipment.

8. Acknowledgments

This work was supported by the National Science Foundation through NSF IIS:1909192 and a gift from Adobe Inc.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[2] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)*, 34(6):1–9, 2015.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[4] Dan A Calian, Jean-François Lalonde, Paulo Gotardo, Tomas Simon, Iain Matthews, and Kenny Mitchell. From faces to outdoor light probes. In *Computer Graphics Forum*, volume 37, pages 51–61. Wiley Online Library, 2018.

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[6] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics*, 39(4), 2020.

[7] Paul Debevec. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia*, 2(4):1–6, 2012.

[8] Paul 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*, pages 145–156, 2000.

[9] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[10] Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. Relighting human locomotion with flowed reflectance fields. In *Rendering Techniques 2006: 17th Eurographics Workshop on Rendering*, pages 183–194, 01 2006.

[11] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[12] Stéphane Grabli, François X Sillion, Stephen R Marschner, and Jerome E Lengyel. Image-based hair capture by inverse lighting. In *Proceedings of Graphics Interface (GI)*, pages 51–58, 2002.

[13] Tim Hawkins, Per Einarsson, and Paul E Debevec. A dual light stage. *Rendering Techniques*, 5(91-98):2, 2005.

[14] Erik Hjelmås and Boon Kee Low. Face detection: A survey. *Computer vision and image understanding*, 83(3):236–274, 2001.

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[16] Yoshihiro Kanamori and Yuki Endo. Relighting humans: Occlusion-aware inverse rendering for full-body human images. *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018*, 37(6), 2018.

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[18] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018.

[19] Chloe Legendre, Wan Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. Learning Illumination from Diverse Portraits. *SIGGRAPH Asia 2020 Technical Communications, SA 2020*, 2020.

[20] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind Video Temporal Consistency via Deep Video Prior. *NeurIPS*, pages 1–11, 2020.

[21] Andrew Liu, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros, and Noah Snavely. Learning to Factorize and Relight a City. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12349 LNCS:544–561, 2020.

[22] Bin Liu, Kun Xu, and Ralph R. Martin. Static Scene Illumination Estimation from Videos with Applications. *Journal of Computer Science and Technology*, 32(3):430–442, 2017.

[23] Xiaoming Liu. Video-based face model fitting using adaptive active appearance model. *Image and Vision Computing*, 28(7):1162–1172, 2010.

[24] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[26] Pieter Peers, Tim Hawkins, and Paul Debevec. A reflective light stage. 2006.

[27] Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. Post-production facial performance relighting using reflectance transfer. *ACM Transactions on Graphics (TOG)*, 26(3):52–es, 2007.

[28] Rolandos Alexandros Potamias, Jiali Zheng, Stylianos Ploumpis, Giorgos Bouritsas, Evangelos Ververas, and Stefanos Zafeiriou. Learning to generate customized dynamic 3d facial expressions. In *European Conference on Computer Vision*, pages 278–294. Springer, 2020.

[29] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: An open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

[30] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214*, 2015.

[31] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing

with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5541–5550, 2017.

[32] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] 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, 2019.

[34] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):157–171, 2019.

[35] O. Wang, J. Davis, E. Chuang, I. Rickard, K. de Mesa, and C. Dave. Video relighting using infrared illumination. *Computer Graphics Forum (Proceedings Eurographics 2008)*, 27(2), Apr. 2008.

[36] Ting Chun Wang, Arun Mallya, and Ming Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *arXiv*, 2020.

[37] Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2008.

[38] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics*, 39(6), 2020.

[39] Zhen Wen, Zicheng Liu, and Thomas S Huang. Face relighting with radiance environment maps. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–158. IEEE, 2003.

[40] Hongyu Wu, Xiaowu Chen, Mengxia Yang, and Zhihong Fang. Facial performance illumination transfer from a single video using interpolation in non-skin region. *Computer Animation and Virtual Worlds*, 24(3-4):255–263, 2013.

[41] Lingbo Yang, Zhanning Gao, Peiran Ren, Siwei Ma, and Wen Gao. Intrinsic temporal regularization for high-resolution human video synthesis. *arXiv*, pages 1–10, 2020.

[42] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[43] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans Peter Seidel, Christian Theobalt, and William A.P. Smith. Self-supervised Outdoor Scene Relighting. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12367 LNCS, pages 84–101, 2020.

[44] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. Deep Image Compositing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 365–374, 2021.

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018.

[47] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020.

[48] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7194–7202, 2019.

[49] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[50] Zhengxia Zou. Castle in the sky: Dynamic sky replacement and harmonization in videos. In *arXiv preprint arXiv:2010.11800*, 2020.