Neural Video Portrait Relighting in Real-time via Consistency Modeling

Longwen Zhang[1,2] *   Qixuan Zhang[1,2] *   Minye Wu[1,3]   Jingyi Yu[1]   Lan Xu[1]

[1]ShanghaiTech University   [2]Deemos Technology
[3]University of Chinese Academy of Sciences

{zhanglw2,zhangqx1,wumy,yujingyi,xulan1}@shanghaitech.edu.cn

{zhanglw,zhangqx}@deemos.com

Abstract

Video portraits relighting is critical in user-facing human photography, especially for immersive VR/AR experience. Recent advances still fail to recover consistent relit result under dynamic illuminations from monocular RGB stream, suffering from the lack of video consistency supervision. In this paper, we propose a neural approach for real-time, high-quality and coherent video portrait relighting, which jointly models the semantic, temporal and lighting consistency using a new dynamic OLAT dataset. We propose a hybrid structure and lighting disentanglement in an encoder-decoder architecture, which combines a multi-task and adversarial training strategy for semantic-aware consistency modeling. We adopt a temporal modeling scheme via flow-based supervision to encode the conjugated temporal consistency in a cross manner. We also propose a lighting sampling strategy to model the illumination consistency and mutation for natural portrait light manipulation in real-world. Extensive experiments demonstrate the effectiveness of our approach for consistent video portrait light-editing and relighting, even using mobile computing.

1. Introduction

The past ten years have witnessed a rapid development of digital portrait photography with the rise of mobile cameras. Relighting evolves as a cutting-edge technique in such portrait photography for immersive visual effects of VR/AR experience. How to further enable consistent relit video results under challenging dynamic illumination conditions conveniently remains unsolved and has received substantive attention in both industry and academia.

For video portrait relighting, early solutions [14, 57] rely on a sophisticated studio setup which is expensive and difficult to be deployed. Modern approaches [41, 44, 42] further

Figure 1. Our approach achieves high-quality and consistent video portrait relighting under dynamic illuminations in real-time, using only mobile computing and monocular RGB video input.

apply color or style transfer techniques to ease the hardware requirements. However, they still require two to four orders of magnitude more time than is available for interactive video application. The recent learning techniques bring huge potential for human portrait modeling and relighting [69, 67, 46, 31, 56] from only monocular RGB input. In particular, the methods [43, 39] perform explicit neural inverse rendering but are limited to the low-quality face and Spherical Harmonics (SH) lighting models. Recent methods [67, 46] remove the explicit inverse rendering by learning an efficient end-to-end mapping between the input headshots and relit ones, while the method [56] further models reflectance attributes explicitly to handle lighting effects like specular or shadow. However, they still focus on single image input without modeling the temporal consistency for video portrait relighting, leading to severe jittery artifacts, especially under the challenging dynamic illuminations. Some recent mobile devices [2, 18] enables the "Portrait Lighting" mode for video editing of lighting conditions. Critically, they only modify existing illumination rather than relight the captured video into various scenes.

In this paper, we address the above challenges and present a novel real-time and temporally coherent portrait relighting approach from only a monocular RGB video input, as illustrated in Fig. 1. Our approach jointly models the semantic, temporal and lighting consistency to enable realistic video portrait light-editing and relighting into new scenes with dynamic illuminations, whilst maintaining real-time performance even on portable device.

Generating such realistic and consistent video relit result in a real-time and data-driven manner is non-trivial. From the data aspect, existing face datasets [10, 60, 56] lack the video ground-truth supervision for consistency modeling. Thus, we build up a high-quality dataset for video portrait relighting, consisting of 603,288 temporal OLAT (one light at a time) images at 25 frames per second (fps) of 36 actors and 2,810 environment lighting maps. From the algorithm side, we further propose a novel neural scheme for consistent video portrait relighting under dynamic illuminations. To maintain the real-time performance, we adopt the encoder-decoder architecture to each input portrait image similar to previous methods [46, 56]. Differently, we introduce a hybrid and explicit disentanglement for semantic-aware consistency, which self-supervises the portrait structure information and fully supervises the lighting information simultaneously in the bottleneck of the network. Such disentanglement is further enhanced via multi-task training as well as an adversarial strategy so as to encode the semantic supervision and enable more realistic relighting. Then, to utilize the rich temporal consistency in our dynamic OLAT dataset, a novel temporal modeling scheme is adopted between two adjacent input frames. Our temporal scheme encodes the conjugated temporal consistency in a cross manner via flow-based supervision so as to model the dynamic relit effect. Finally, a lighting sampling based on Beta distribution is adopted, which augments the discrete environment lighting maps and generates a triplet of lighting conditions for adjacent input frames and the target output. Our sampling scheme models the illumination consistency and mutation simultaneously for natural video portrait light-editing and relighting in the real-world. To summarize, our main contributions include:

- We present a real-time neural video portrait relighting approach, which faithfully models the video consistency for dynamic illuminations, achieving significant superiority to the existing state-of-the-art.

- We propose an explicit structure and lighting disentanglement, a temporal modeling as well as a lighting sampling schemes to enable realistic video portrait light-editing and relighting on-the-fly.

- We make available our dataset with 603,288 temporal OLAT images to stimulate further research of human portrait and lighting analysis.

2. Related Work

Portrait Relighting. Debevec *et al.* [14] invented Light Stage to capture the reflectance field of human faces, which has enabled high-quality 3D face reconstruction and illuminations rendering, advancing the film's special effects industry. Some subsequent work has also achieved excellent results by introducing deep learning [64, 29, 21, 30, 47, 49]. Obviously, this is not a product for individual consumers; thus various methods for single portrait relighting have been proposed. Several methods [58, 6, 26, 59, 50] perform relighting on static objects. Some work follows the pipeline of color transfer to achieve the relighting effects [11, 41, 44, 42], which usually needs another portrait image as the facial color distribution reference. Blanz *et al.* [7] use a morphable model of faces that allows relighting by changing the directional lighting model parameters. Several work [67, 28, 39, 1, 15, 54] estimate SH [5, 34] parameters and achieve relighting by modifying the parameters of the SH lighting model. Sevastopolsky *et al.* [40] use point cloud to generate relightable 3D head portraits, while Tewari *et al.* [48] use GAN to manipulate the illumination. Explicitly modeling the shadow and specular [56, 31] achieve excellent results in directional light source relighting. Mallikarjunr *et al.* [4] take a single image portrait as input to predict OLAT as Reflectance Fields, which can be relit to other lighting via image-based rendering. Sun *et al.* [46] choose environment map as lighting model and use light stage captured OLAT data to generate realistic training data and train relighting networks in an end-to-end fashion. We also use the OLAT images to generate training data for portrait relighting. Differently, our approach enables real-time and consistent video portrait relighting under dynamic illuminations, with the aid of a new dynamic OLAT dataset.

Temporal Consistency. Previous image relighting methods can be extended to videos if we directly treat every video frames as independent images. However, those methods will inevitably generate flicker results on relit videos. To suppress flicker results, several approaches have been developed for video style transfer tasks [27, 23, 8, 61, 51, 9, 36, 53]. Specifically, Ruder *et al.* [36] employed a temporal loss guided by optical flow for video style transfer, but the real-time computation of optical flows makes this approach slower. Vid2Vid [53] synthesised videos with temporal consistency by training a network to estimate optical flow and apply it on previously generated frames. In this paper, we show that temporal consistency and portrait relighting can be simultaneously learned by a feed-forward CNN, which avoids computing optical flows in the inference stage.

Video Relighting. Some methods [52, 20, 37, 55] use LEDs or OLAT to provide acceptable lighting for video conferencing. ST-NeRF [63] manipulates dynamic scene through a spatio-temporal coherent neural layered radiance
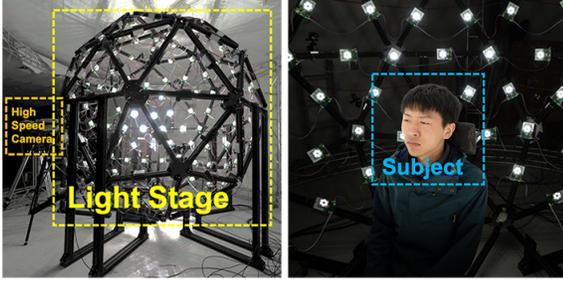
Figure 2. Illustration of the capturing system for building our dynamic OLAT dataset.

Figure 3. Illustration of our video portrait relighting network based on encoder-decoder architecture for real-time inference.

representation. Li *et al.* [25] create free-viewpoint relighting video using multi-view reconstruction under general illumination, while Richardt *et al.* [35] add video effects like relighting using RGBZ video cameras. "The Relightables" [21] proposes a hybrid geometric and machine learning reconstruction pipeline to generate high-quality relit video. In contrast, our method does not require extra specific capturing equipment and enables real-time video portrait relighting using mobile computing.

Face Dataset. Traditional face dataset usually takes 2D images in various lighting conditions [19, 24, 16]. The controlled lighting conditions are easy to build but lack reflectance information for photo-realistic portrait relighting. With the development of face scanning and reconstruction techniques, 3D face datasets have been extended from only geometric [62, 65, 38, 10, 12, 13, 66, 60] to include reflectance channels [45, 56]. However, the existing rendering scheme is difficult to avoid the uncanny valley effect without manual modification. 3D dataset still cannot achieve the realism in 2D face dataset or those using image-based rendering Thus, various face OLAT datasets [4, 46, 31] have been proposed with light stage setup. In contrast, we construct a new dynamic OLAT dataset through a a light stage setup and a 4K ultra-high-speed camera. Our high-quality dataset consists of 603,288 temporal OLAT imagesets at 25 fps of 36 subjects (18 females and 18 males) with various expressions and hairstyles.

3. Dynamic OLAT Dataset Overview

Our goal is to naturally manipulate the lighting of a portrait RGB video captured in the wild into new environment lighting conditions while preserving the consistent structure and content. To provide ground truth supervision for video consistency modeling, we build up a high-quality dynamic OLAT dataset. As illustrated in Fig. 2, our capture system consists of a light stage setup with 114 LED light sources and Phantom Flex4K-GS camera (global shutter, stationary 4K ultra-high-speed camera at 1000 fps), resulting in dynamic OLAT imageset recording at 25 fps using the overlapping method [57]. Our dataset includes 603,288 temporal OLAT imagesets of 36 actors (18 females and 18 males)
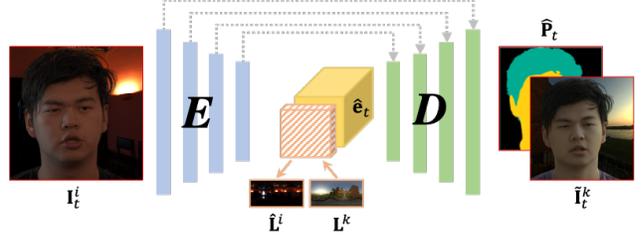
with 2810 HDR environment lighting maps [22, 17], to provide high-quality and diverse video portrait relighting samples. Besides, we apply a pre-defined light condition to each OLAT imageset to obtain a fully-illuminated portrait image. Then, both the portrait parsing [68] and the optical flow [32] algorithms are applied to such a fully-illuminated stream to obtain ground truth semantics and correspondence supervisions. Our dynamic OLAT dataset provides sufficient semantic, temporal and lighting consistency supervision to train our neural video portrait relighting scheme, which can generalize to in-the-wild scenarios.

4. Neural Video Portrait Relighting

Our scheme obtains real-time and coherent video relit results from only mobile computing and monocular RGB steam, using our dynamic OLAT dataset, as shown in Fig. 4. First, a hybrid disentanglement for portrait structure and lighting is introduced for semantic-aware consistency modeling (Sec. 4.1). Then, a temporal modeling is adopted to utilize the flow-based supervision for dynamic relit effects (Sec. 4.2). We also introduce a lighting sampling strategy to model the illumination consistency and mutation for natural portrait lighting manipulation in real-world scenarios (Sec. 4.3).

Notations. To achieve the real-time performance, we adopt the encoder-decoder architecture to the portrait stream sequentially similar to previous methods [46, 56], as illustrated in Fig. 3. Given an input image $\mathbf{I}_t^i$ at t-th frame and a desired target lighting $\mathbf{L}^k$ (environment lighting map), our network Φ predicts both a target portrait image $\tilde{\mathbf{I}}_t^k$ lit by the $\mathbf{L}^k$ and a corresponding semantic mask $\hat{\mathbf{P}}_t$:

$$\tilde{\mathbf{I}}_t^k, \hat{\mathbf{L}}^i, \hat{\mathbf{P}}_t = \Phi(\mathbf{I}_t^i, \mathbf{L}^k), \qquad (1)$$
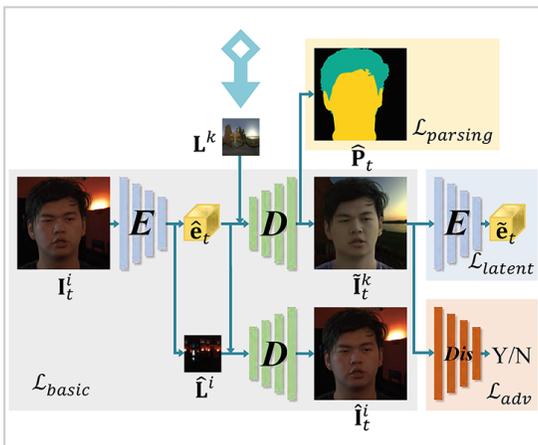
where $\hat{\mathbf{L}}^i$ is the regressed lighting of the input image $\mathbf{I}_t^i$. Specifically, the encoder Φ_{enc} encodes the input image into both the lighting $\mathbf{L}^i$ and a portrait structure latent code $\hat{\mathbf{e}}_t$:

$$\hat{\mathbf{L}}^i, \mathbf{F}_t^i, \hat{\mathbf{e}}_t = \Phi_{enc}(\mathbf{I}_t^i), \qquad (2)$$
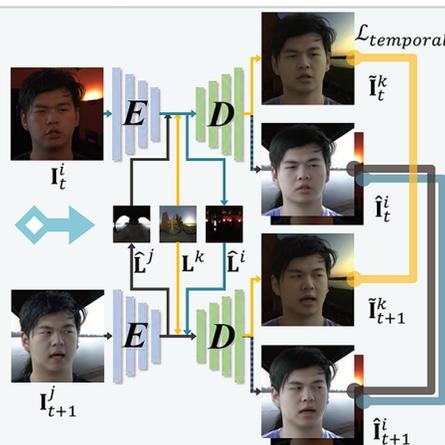
where $\mathbf{F}_t^i$ is the output of skip connections. Similarly, the corresponding decoder Φ_{dec} is formulated as:

$$\hat{\mathbf{I}}_t^i, \hat{\mathbf{P}}_t = \Phi_{dec}(\hat{\mathbf{L}}^i, \mathbf{F}_t^i, \hat{\mathbf{e}}_t), \qquad (3)$$
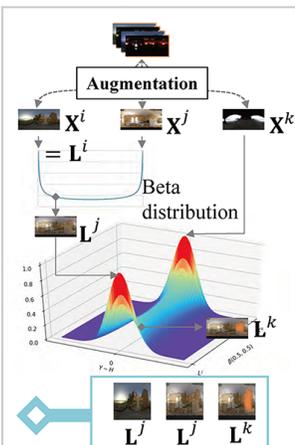
Figure 4. The training pipeline of our approach. It consists of a structure and lighting disentanglement (Sec. 4.1), a temporal consistency modeling (Sec. 4.2) and a lighting sampling (Sec. 4.3), so as to generate consistent video relit results from a RGB stream in real-time.

where we use predicted illumination $\hat{\mathbf{L}}^i$ to relit itself. By replacing the lighting with a known one $\mathbf{L}^k$ in the decoder, we can obtain a relit portrait image $\tilde{\mathbf{I}}_t^k$ corresponding to $\mathbf{L}^k$. Note that we use tilde and hat notations for the images relit by known illumination or the predicted one, respectively.

4.1. Structure and Lighting Disentanglement

The core of realistic relighting is the reliable disentanglement of the portrait structure information $\mathbf{e}$ and the lighting condition $\mathbf{L}$ in the bottleneck of our network in Eqn. 2. To this end, we adopt a hybrid disentanglement scheme combing with multi-task training and adversarial training strategy to model the semantic consistency for realistic relighting.

Similar to method [46], during training we optimize the following basic loss for disentanglement, which minimizes the photometric error and illumination distance between the prediction and the ground truth from our dataset:

$$\mathcal{L}_{basic} = \frac{1}{2} \| \log(1 + \mathbf{L}^i) - \log(1 + \hat{\mathbf{L}}^i) \|_2^2 \\ + \| \mathbf{M}_t \odot (\mathbf{I}_t^k - \tilde{\mathbf{I}}_t^k) \|_1 + \| \mathbf{M}_t \odot (\mathbf{I}_t^i - \hat{\mathbf{I}}_t^i) \|_1, \tag{4}$$

where $\mathbf{M}_t$ is a portrait foreground mask of the t corresponding frame from face parsing and $\odot$ is element-wise multiplication. However, only using this basic scheme fails to provide supervision on the structure latent code $\mathbf{e}$ and encode the rich semantic consistency in our dynamic OLAT dataset. Thus, we introduce the following strategies for more reliable disentanglement and more consistent relighting.

Structure Self-supervision. Recall that the method [46] treats the latent space as lighting map and only relies on the feature maps $\mathbf{F}_t^i$ for modeling the portrait structure information. Differently, we utilize a separated structure-wise latent code $\hat{\mathbf{e}}_t$, which has larger receptive field in the encoder to represent the global context information of the portrait im-

age. Thus, we design a novel self-supervised scheme for such structure-wise latent code by applying the encoder recurrently to the relit output and enforces the consistency between the structure codes for further disentanglement. We formulate it as follows:

$$\mathcal{L}_{latent} = \| \hat{\mathbf{e}}_t - \tilde{\mathbf{e}}_t \|_2^2, \tag{5}$$

where $\tilde{\mathbf{e}}_t$ is from the recurrent output of $\Phi_{enc}(\tilde{\mathbf{I}}_t^k)$. Here we utilize encoder Φ_{enc} to encode the relit image $\tilde{\mathbf{I}}_t^k$ with the target light $\mathbf{L}^k$ and verify its global structure latent code with the original one to enhance structure consistency.

Semantics-aware Multi-Task Learning. Human face skin with scattering effects owns different reflectance distribution from other materials like hair. Treating all the portrait pixels uniformly will cause strong artifacts in relighted results without maintaining the semantic consistency. To this end, we design the multi-task decoder Φ_{dec} which aims to restore both the relighted image $\hat{\mathbf{I}}_t^o$ and the semantic portrait mask $\mathbf{P}_t$ under the supervision from our dataset. Such parsing loss with binary cross-entropy metrics is formulated as:

$$\mathcal{L}_{parsing} = -\Big(\mathbf{P}_t \odot \log \hat{\mathbf{P}}_t + (1 - \mathbf{P}_t) \odot \log(1 - \hat{\mathbf{P}}_t) \Big). \tag{6}$$

By predicting semantic portrait mask $\mathbf{P}_t$, we enforce both encoder and decoder networks to be aware of the semantic information of human portrait image. Hence the networks can implicitly model the semantic consistency for more realistic relighting of various portrait regions.

Adversarial Training. To further reinforcing portrait image details, we also introduce a discriminator network Φ_D, which has the same architecture as DCGAN's [33]. We adopt Wasserstein GAN [3] strategies to the proposed discriminator for a stable training process. Specifically, we

remove the activation function of Φ_D's final layer and apply the weight clipping method during training. Then, the adversarial losses is formulated as:

$$\begin{aligned} \mathcal{L}_{adv_D} &= -\Phi_D(\mathbf{I}_t^i, \mathbf{I}_t^k, \mathbf{L}^k) + \Phi_D(\mathbf{I}_t^i, \tilde{\mathbf{I}}_t^k, \mathbf{L}^k) \\ \mathcal{L}_{adv_G} &= -\Phi_D(\mathbf{I}_t^i, \tilde{\mathbf{I}}_t^k, \mathbf{L}^k), \end{aligned} \quad (7)$$

where $\mathcal{L}_{adv_D}$ is only for updating the discriminator, and $\mathcal{L}_{adv_G}$ is only for updating the decoder; $\mathcal{L}_{adv} = \mathcal{L}_{adv_D} + \mathcal{L}_{adv_G}$. Here, the discriminator takes a triplet as input, including a source image, a relit image, and the corresponding light condition $\mathbf{L}^k$, which estimates the Wasserstein distance between the real image distribution and the relit image distribution. Note that the structure of the source image and $\mathbf{L}^k$ are essential cues for such distance measurement.

4.2. Temporal Consistency Enhancement

Previous single-image relighting approaches lack the explicit temporal modeling in terms of portrait motions or dynamic illuminations, leading to flickering artifacts for video applications. Thus, we propose a rescue by utilizing the rich temporal consistency supervision in our dynamic OLAT dataset. Specifically, for two adjacent training samples at timestamps t and $t+1$, we obtain the forward flow $f_{t,t+1}(\cdot)$ and the backward flow $f_{t+1,t}(\cdot)$ from our continues OLAT imagesets, where $f_{a,b}(\cdot)$ warps images from time a to time b. Note that our high-quality OLAT imagesets at 25 fps ensures the accuracy of such supervision based on optical flow for self-supervised verification in a cross manner.

For balancing the lighting distribution between two frames, we also introduce two conjugated illuminations in our training scheme. Given two adjacent frames $\mathbf{I}_t^i$ and $\mathbf{I}_{t+1}^j$, we relit both of them using our network according to the predicted lighting conditions $\hat{\mathbf{L}}^i, \hat{\mathbf{L}}^j$, as well as the target illumination $\mathbf{L}^k$ to obtain corresponding relit images at both frames. Thus, our temporal loss is formulated as:

$$\begin{aligned} \mathcal{L}_{temporal} &= \|f_{t,t+1}(\tilde{\mathbf{I}}_t^k) - \tilde{\mathbf{I}}_{t+1}^k\|_1 + \|f_{t+1,t}(\tilde{\mathbf{I}}_{t+1}^k) - \tilde{\mathbf{I}}_t^k\|_1 \\ &+ \sum_{z \in \{i,j\}} \left(\|f_{t,t+1}(\hat{\mathbf{I}}_t^z) - \hat{\mathbf{I}}_{t+1}^z\|_1 + \|f_{t+1,t}(\hat{\mathbf{I}}_{t+1}^z) - \hat{\mathbf{I}}_t^z\|_1 \right), \end{aligned}$$
$$(8)$$

which encodes the conjugated temporal consistency in a cross manner via flow-based supervision so as to model the dynamic relit effect.

4.3. Lighting Conditions Sampling

Note that the discrete environment lighting maps in our dynamic OLAT dataset still cannot model the illumination consistency and mutation for real-time video relighting scenarios. Thus, we introduce a novel lighting sampling scheme during training to generate a triplet of lighting conditions for adjacent input frames and the target output, which enhances the illumination consistency for natural portrait relighting and lighting manipulation.

Similar to previous work [46, 56], the lighting condition $\mathbf{L}$ is represented as a flattened latitude-longitude format of size 16×16 with three color channels from the environment lighting map and we adopt the same environment map re-rendering with random rotation to augment the lighting conditions in our dataset, which forms a uniform lighting condition sampling distribution $\mathcal{G}$. To model illumination mutation, we further design a lighting condition distribution $\mathcal{H}$, where we randomly sample one to three point light sources with uniformly random colors outside a unit sphere. The maximum distance from the light source to the sphere is limited by 1.5 in order to produce reasonable illumination. Then, we project these light sources on a sphere according to the Lambertian reflectance model [5] to form an environment map and corresponding lighting condition.

For each training sample, we generate three illuminations using different sampling strategies, including $\mathbf{L}^i$ and $\mathbf{L}^j$ to generate the adjacent images $\mathbf{I}_t^i$ and $\mathbf{I}_{t+1}^j$, as well as the $\mathbf{L}^k$ for the target image $\mathbf{I}^k$, which is formulated as:

$$\begin{aligned} \mathbf{L}^i &= \mathbf{X}^i \\ \mathbf{L}^j &= \beta_1 \mathbf{L}^i + (1 - \beta_1)\mathbf{X}^j \\ \mathbf{L}^k &= \beta_2 \mathbf{L}^j + (1 - \beta_2)\mathbf{X}^k + \mathbf{Y}, \end{aligned} \quad (9)$$

where $\mathbf{X}^i, \mathbf{X}^j, \mathbf{X}^k \sim \mathcal{G}$, $\mathbf{Y} \sim \mathcal{H}$ and β_1, β_2 are sampled from a Beta distribution $Beta(0.5, 0.5)$. Here, the Beta distribution drastically diversifies the lighting combination for modeling the illumination consistency and mutation simultaneously and enhancing the generalization ability of our network. Conceptually, $\mathbf{L}^i$ and $\mathbf{L}^j$ have similarity, contributing to the coverage of temporal loss training. And $\mathbf{L}^k$ provides challenging illumination examples to enhance the lighting manipulation ability of our approach.

4.4. Implementation Details

We utilize our dynamic OLAT dataset to train our video relighting network. Our training dataset consists of 536,256 temporal OLAT imagesets of 32 actors lasting for 188.16 seconds. The remaining OLAT imagesets of the other four actors unseen during training are taken as the test dataset. Note that we also augment the dataset using random cropping and resizing to add more challenging incomplete portraits to enhance the generalization ability of our network. During training our total loss is formulated as follows:

$$\begin{aligned} \mathcal{L} = \lambda_1 \mathcal{L}_{basic} &+ \lambda_2 \mathcal{L}_{latent} + \lambda_3 \mathcal{L}_{parsing} \\ &+ \lambda_4 \mathcal{L}_{temporal} + \lambda_5 \mathcal{L}_{adv}, \end{aligned} \quad (10)$$

where the weights for each term are set to be 1.0 in our experiments. Since we utilize Wasserstein GAN in our approach, our network's parameters are optimized by RMSprop algorithm with a learning rate of 0.00005. Besides, we also clamp parameters of discriminator into a range of $[-0.01, 0.01]$ and adopt the progressive training strategy.
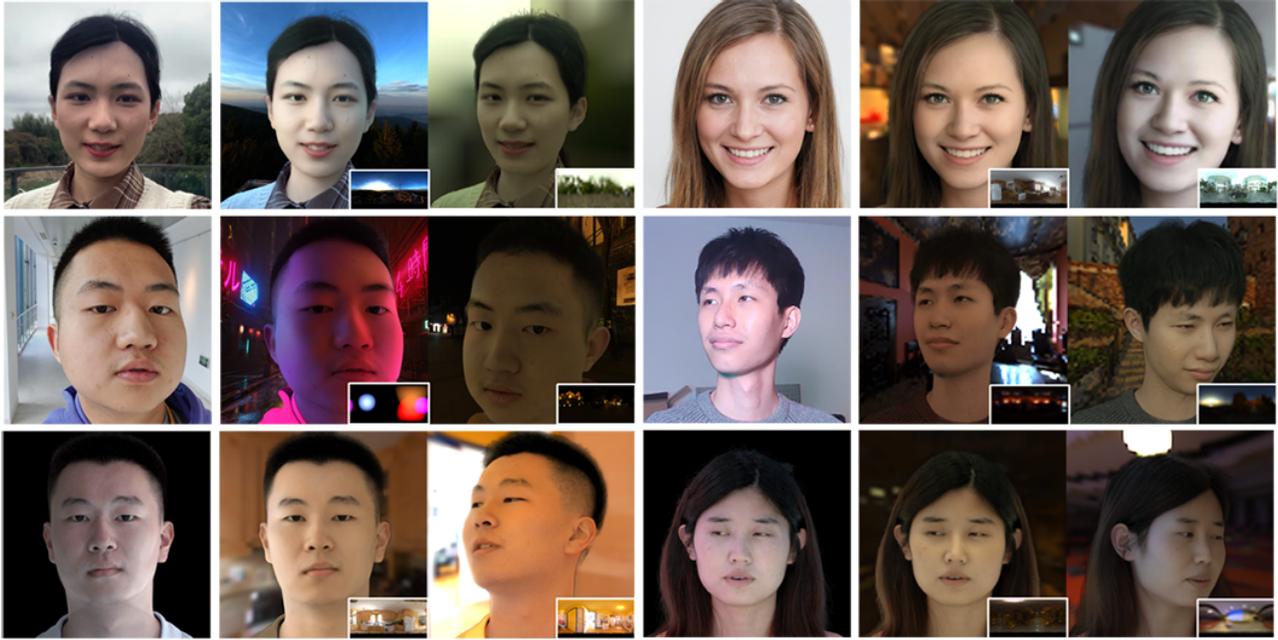
Figure 5. Our relighting results under dynamic illuminations. Each triplet includes the input frame and two relit result examples.

| Input | Ground Truth | Ours | SIPR | EMRCM | MTP | DPR |

Figure 6. Qualitative comparisons of relit results on our dynamic OLAT dataset. Our approach achieves more realistic relighting. The small images in MTP and DPR are the reference and the gray-scale SH lighting, respectively. Please refer to [42, 67] for details.

5. Experiments

Here we evaluate our method in various challenging scenarios. We run our experiments on a PC with Intel i7-8700K CPU and Nvidia RTX 3080 GPU, where our approach generates high-quality 512×512 relit results at 111 fps (15 fps on iPhone 12) with the latency of 29.2 ms. Fig. 5 demonstrates several results of our approach, which can generate consistent video relit results of both in-the-wild sequences and the one from our dataset with challenging illuminations.

5.1. Comparison

We compare our approach against existing state-of-the-art methods, including Single Image Portrait Relighting (**SIPR**) [46], the one via Explicit Multiple Reflectance Channel Modeling (**EMRCM**) [56], the one based on Mass Transport Approach (**MTP**) [42] and Deep Portrait Relighting (**DPR**) [67]. Note that we re-implement SIPR [46] and train it using our dataset for a fair comparison. Please refer to supplementary material for the results from the original SIPR author. Fig. 6 and Fig. 7 provide the qualitative comparison on both our dynamic OLAT dataset and online or in-the-wild sequences, respectively. Note that our approach achieves significantly more realistic relit results under challenging illuminations by modeling the video consistency.

Then, we utilize our testing set with ground truth for quantitative comparison. Similar to previous methods [46, 56], we adopt the **RMSE**, **PSNR**, and **SSIM** as metrics. Note that the output values are normalized to $[0, 1]$, and only the valid portrait regions are considered. As shown

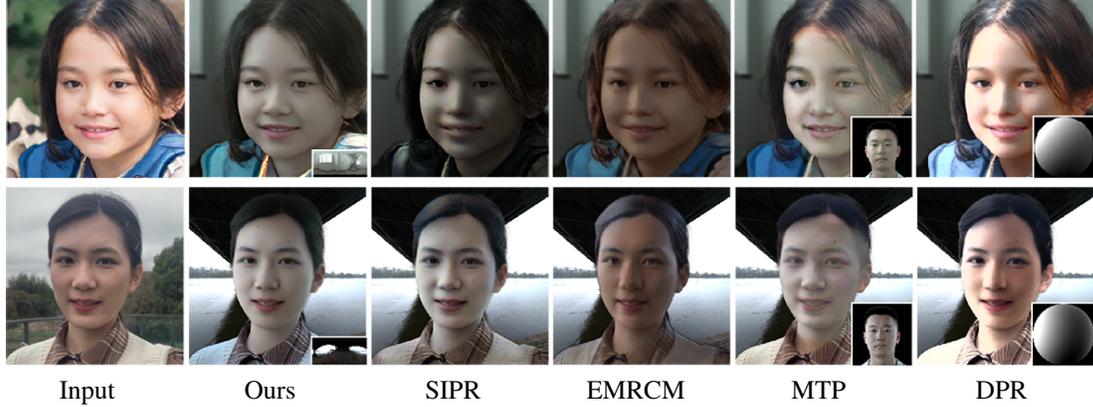| Input | Ours | SIPR | EMRCM | MTP | DPR |

Figure 7. Qualitative comparisons of relit results on online or in-the-wild sequences. We obtain more natural results on unseen performers.
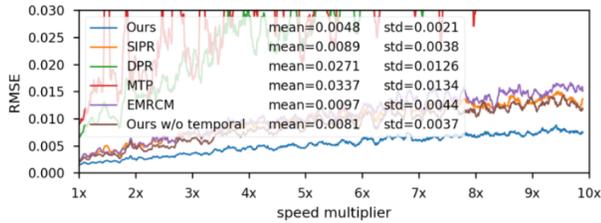

Figure 8. Quantitative comparison in terms of handling fast changing illuminations. Our approach consistently outperforms other baselines under various lighting speed up factors.

Method	RMSE	PSNR	SSIM
SIPR	0.0974	20.6542	0.8901
EMRCM	0.0766	22.7197	0.8748
MTP	0.0902	21.9535	0.8775
DPR	0.1080	20.8042	0.8593
Ours	**0.0349**	**30.6110**	**0.9584**

Table 1. Quantitative comparison on our dynamic OLAT dataset.

in Tab. 1, our approach consistently outperforms the baselines in terms of these metrics above. We further compare against baselines under dynamic illuminations. Thus, we synthesize 1000 frames with the static performer and changing lighting conditions using various speed-up factors from 1 to 10. Then, we relit the sequence into static lighting and calculated the average RMSE of adjacent output frames as the error metric for the jittery artifacts. As shown in Fig. 8, the error of our approach glows much slower consistently compared to others, which illustrates our superiority to handle dynamic illuminations.

5.2. Evaluation

Hybrid disentanglement. Here we evaluate our hybrid scheme for structure and lighting disentanglement. Let **w/o structure** denote the variation of our approach without the self-supervision of portrait structure in Eqn. 5, and **w/o enhance** denote the variation without the enhance-
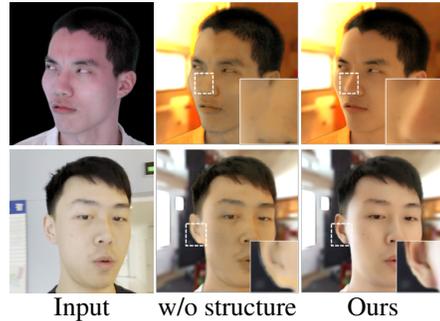

| Input | w/o structure | Ours |

Figure 9. Qualitative evaluation of structure self-supervison. Our full pipeline achieve sharper relighting with fine structured details.
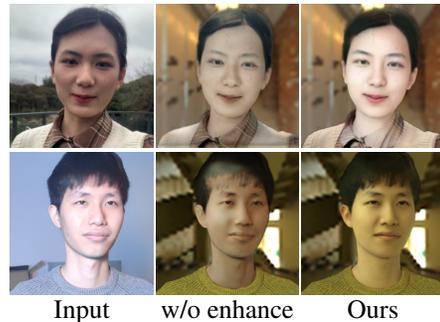

| Input | w/o enhance | Ours |

Figure 10. Qualitative evaluation of disentangle enhancement. Our scheme models the semantic consistency for realistic relighting.

ment of disentanglement by using multi-task and adversarial strategy. As shown in Fig. 9, our scheme with structure self-supervision enables more accurate disentanglement for sharper realistic results. The qualitative evaluation in Fig. 10 further illustrates that our multi-task and adversarial training strategy encodes the semantic consistency for more realistic relighting.

Temporal modeling. Here we compare against our variation without temporal consistency modeling, denoted as **w/o temporal**. Similar to the comparison under dynamic illuminations, we relit the same synthesized sequence with static performer and changing illuminations into a target
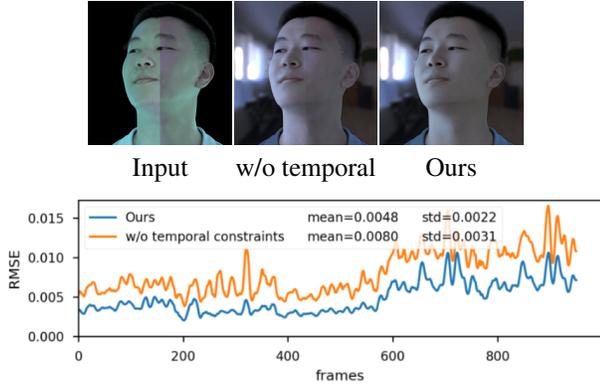
Figure 11. Evaluation of temporal modeling. Top: two relit examples of a static portrait from changing illuminations into a target lighting condition. Down: the corresponding error curve.

Method	RMSE	PSNR	SSIM
w/o content	0.0549	25.5496	0.9021
w/o temporal	0.0404	28.6403	0.9510
w/o parsing	0.0680	23.6189	0.9170
w/o sampling	0.0616	24.5142	0.9223
Ours	**0.0349**	**30.6110**	**0.9584**

Table 2. Quantitative evaluation on synthetic sequences.

lighting condition for thorough evaluation. As shown in Fig. 11, our approach with temporal modeling achieves more temporal consistent results both qualitatively and quantitatively. We also provide quantitative evaluation under various lighting speed-up factors in Fig. 8, which illustrates the effectiveness of utilizing temporal consistency.

Lighting Sampling We further evaluate our light sampling strategy. Let **w/o sampling** denote our variation only using the discrete environment lighting maps during training. As shown in Fig. 12 and Fig. 13, our scheme models the illumination consistency and mutation, enabling a more natural portrait light-editing and relighting.

We further perform thorough quantitative analysis of the individual components of our approach using our testing set with ground truth. As shown in Tab. 2, our full pipeline consistently outperforms other variations.

6. Discussion and Conclusion

Limitations. As a trial to explore real-time and consistent video portrait relighting under dynamic illuminations, our approach still has a number of limitations as follows. First, our approach cannot handle the extreme illumination changes, like suddenly turn on/off all the lights. Besides, the generated relit video results lose the facial details partially due to the light-weight encoder-decoder architecture. Also, the relit results have inaccurate specularities and are often over-diffused. The relit videos sometimes have temporal artifacts near the hair boundaries due to the inconsistent portrait matting during inference. Our current approach
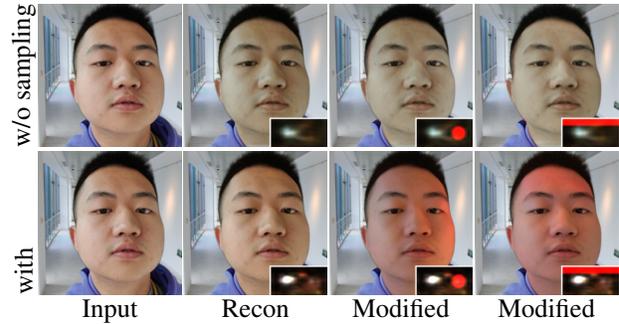

Figure 12. Evaluation of lighting sampling on real-world scenes. Our scheme enables more natural lighting manipulation, where a red light source is added into the environment lighting map.
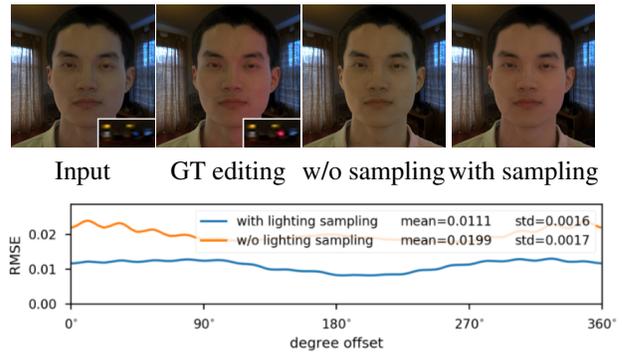

Figure 13. Evaluation of lighting sampling on synthetic scenes where a red light is added from different directions. Top: the qualitative relit examples. Down: the corresponding error curve.

is also limited to headshots only. It's promising to include clothes and garment material analysis for full-body portrait relighting. It's also interesting to enhance the OLAT dataset with a generative model to handle high-frequency lighting.

Conclusion. We have presented a novel scheme for real-time, high-quality and consistent video portrait relighting under dynamic illuminations from monocular RGB stream and a new dynamic OLAT dataset. Our hybrid disentangles scheme with a multi-task and adversarial training models the semantic consistency efficiently and generates realistic relit results. Our temporal modeling scheme encodes the flow-based supervision for temporally consistent relighting, while our light sampling strategy enhances the illumination consistency for lighting manipulation. We believe that our approach is a critical step for portrait lighting analysis, with many potential applications in user-facing photography, VR/AR visual effects or immersive telepresence.

Acknowledgments

References

[1] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3d morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1080–1093, 2012. 2

[2] Apple. Use portrait mode on your iphone. https://support.apple.com/en-us/HT208118. 1

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 4

[4] Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Monocular reconstruction of neural face reflectance fields. 2, 3

[5] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 2, 5

[6] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[7] 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

[8] Nicolas Bonneel, Kalyan Sunkavalli, Sylvain Paris, and Hanspeter Pfister. Example-based video color grading. *ACM Transactions on Graphics*, 32, 2013. 2

[9] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. volume 34, 2015. 2

[10] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 2, 3

[11] Xiaowu Chen, Mengmeng Chen, Xin Jin, and Qinping Zhao. Face illumination transfer through edge-preserving filters. In *CVPR 2011*, pages 281–287. IEEE, 2011. 2

[12] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5117–5126, 2018. 3

[13] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 international conference on computer vision*, pages 2296–2303. IEEE, 2011. 3

[14] 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. 1, 2

[15] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 2

[16] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2007. 3

[17] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 36(6):1–14, 2017. 3

[18] Google. Portrait light: Enhancing portrait lighting with machine learning. https://ai.googleblog.com/2020/12/portrait-light-enhancing-portrait.html. 1

[19] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 3

[20] P. Gunawardane, T. Malzbender, R. Samadani, A. McReynolds, D. Gelb, and J. Davis. Invisible light: Using infrared for video conference relighting. In *2010 IEEE International Conference on Image Processing*, pages 4005–4008, 2010. 2

[21] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), Nov. 2019. 2, 3

[22] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6927–6935, 2019. 3

[23] Manuel Lang, Oliver Wang, Tunc Aydin, Aljoscha Smolic, and Markus Gross. Practical temporal consistency for image-based graphics applications. *ACM Transactions on Graphics*, 31, 2012. 2

[24] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005. 3

[25] Guannan Li, Yebin Liu, and Qionghai Dai. Free-viewpoint video relighting from multi-view sequence under general illumination. *Machine vision and applications*, 25(7):1737–1746, 2014. 3

[26] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single

image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2

[27] Peter Litwinowicz. Processing images and video for an impressionist effect. 1997. 2

[28] Yang Liu, Alexandros Neophytou, Sunando Sengupta, and Eric Sommerlade. Relighting images in the wild with a self-supervised siamese auto-encoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–40, 2021. 2

[29] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[30] Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–21, 2020. 2

[31] 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. 1, 2, 3

[32] Nvidia. Nvidia optical flow sdks. `https://developer.nvidia.com/opticalflow-sdk`. 3

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4

[34] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *JOSA A*, 18(10):2448–2459, 2001. 2

[35] Christian Richardt, Carsten Stoll, Neil A Dodgson, Hans-Peter Seidel, and Christian Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of rgbz videos. In *Computer graphics forum*, volume 31, pages 247–256. Wiley Online Library, 2012. 3

[36] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. volume 9796 LNCS, 2016. 2

[37] S. Ryu, S. H. Lee, S. C. Ahn, and J. Park. Tangible video teleconference system using real-time image-based relighting. *IEEE Transactions on Consumer Electronics*, 55(3):1162–1168, 2009. 2

[38] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *European workshop on biometrics and identity management*, pages 47–56. Springer, 2008. 3

[39] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 1, 2

[40] Artem Sevastopolsky, Savva Ignatiev, Gonzalo Ferrer, Evgeny Burnaev, and Victor Lempitsky. Relightable 3d head portraits from a smartphone video. *arXiv preprint arXiv:2012.09963*, 2020. 2

[41] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. 2014. 1, 2

[42] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics*, 37(1), 2017. 1, 2, 6

[43] 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. 1

[44] Yibing Song, Linchao Bao, Shengfeng He, Qingxiong Yang, and Ming-Hsuan Yang. Stylizing face images via multiple exemplars. *Computer Vision and Image Understanding*, 162:135–145, 2017. 1, 2

[45] Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In *Face and Gesture 2011*, pages 611–618. IEEE, 2011. 3

[46] Tiancheng Sun, Jonathan T. Barron, Yun Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics*, 38(4), jul 2019. 1, 2, 3, 4, 5, 6

[47] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)*, 39(6):1–12, 2020. 2

[48] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[49] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020. 2

[50] Jiandong Tian, Zachary Murez, Tong Cui, Zhen Zhang, David Kriegman, and Ravi Ramamoorthi. Depth and image restoration from light field in a scattering medium. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2401–2410, 2017. 2

[51] Peter V., Black Michael J Kong Naejin, and Gehler. Intrinsic video. pages 360–375. Springer International Publishing, 2014. 2

[52] Oliver Wang, James Davis, Erika Chuang, Ian Rickard, Krystle De Mesa, and Chirag Dave. Video relighting using infrared illumination. In *Computer Graphics Forum*, volume 27, pages 271–279. Wiley Online Library, 2008. 2

[53] Ting Chun Wang, Ming Yu Liu, Jun Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. volume 2018-December, 2018. 2

[54] Yang Wang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. Face re-lighting from a single image under harsh lighting conditions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[55] Youjia Wang, Taotao Zhou, Minzhang Li, Teng Xu, Minye Wu, Lan Xu, and Jingyi Yu. Relightable neural video portrait, 2021. 2

[56] 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):1–13, nov 2020. 1, 2, 3, 5, 6

[57] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics (TOG)*, 24(3):756–764, 2005. 1, 3

[58] Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. Deep view synthesis from sparse photometric images. *ACM Transactions on Graphics (TOG)*, 38(4):1–13, 2019. 2

[59] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 2

[60] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 601–610, 2020. 2, 3

[61] Genzhi Ye, Elena Garces, Yebin Liu, Qionghai Dai, and Diego Gutierrez. Intrinsic video and applications. volume 33, 2014. 2

[62] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 3

[63] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Trans. Graph.*, 40(4), July 2021. 2

[64] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. Neural light transport for relighting and view synthesis. *ACM Trans. Graph.*, 40(1), Jan. 2021. 2

[65] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013. 3

[66] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 3

[67] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 7193–7201. Institute of Electrical and Electronics Engineers Inc., oct 2019. 1, 2, 6

[68] zllrunning. face-parsing.pytorch. GitHub repository, 2019. https://github.com/zllrunning/face-parsing.PyTorch. 3

[69] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum (Eurographics State of the Art Reports 2018)*, 37(2), 2018. 1