Real-time 3D-aware Portrait Video Relighting

Om Shrivastava | Aniket Saha | Dr. Tushar Sandhan

Indian Institute of Technology Kanpur, CVPR 2024

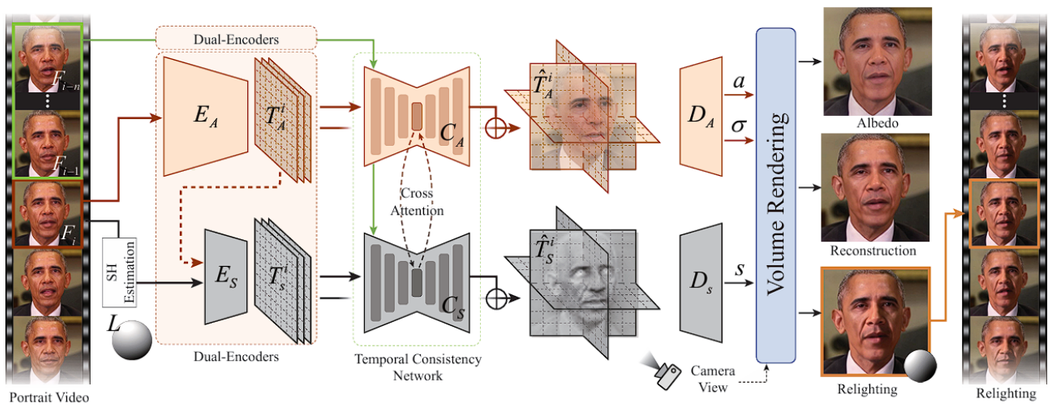
Abstract

Synthesizing realistic videos of talking faces under cus tom lighting conditions and viewing angles benefits various downstream applications like video conferencing. However, most existing relighting methods are either time-consuming or unable to adjust the viewpoints. In this paper, we present the first real-time 3D-aware method for relighting in-the wild videos of talking faces based on Neural Radiance Fields (NeRF). Given an input portrait video, our method can synthesize talking faces under both novel views and novel lighting conditions with a photo-realistic and disen tangled 3D representation. Specifically, we infer an albedo tri-plane, as well as a shading tri-plane based on a de sired lighting condition for each video frame with fast dual encoders. We also leverage a temporal consistency network to ensure smooth transitions and reduce flickering artifacts. Our method runs at 32.98 fps on consumer-level hardware and achieves state-of-the-art results in terms of reconstruc tion quality, lighting error, lighting instability, temporal consistency and inference speed. We demonstrate the effec tiveness and interactivity of our method on various portrait videos with diverse lighting and viewing conditions.

Introduction

Portrait videos have become central to applications across a range of fields, including video conferencing, virtual reality (VR), augmented reality (AR), entertainment, and video editing. Despite their growing use, capturing high-quality portrait videos under diverse lighting conditions remains challenging. Often, videos are recorded in environments with suboptimal lighting or virtual backgrounds that clash with the foreground lighting, which compromises realism and detracts from the user experience. Dynamic relighting of portrait videos to match the virtual or physical environment is particularly valuable in AR and VR applications, where users desire interactive, adaptable 3D models of their faces. Achieving realistic, real-time relighting, however, is complex due to the need to accurately model the interplay between light, facial geometry, and texture. Furthermore, ensuring temporal coherence between video frames without flickering or inconsistencies poses an additional challenge.

Previous methods have attempted to address these issues, but most face significant limitations. Many approaches can only relight from fixed viewpoints, restricting the user's freedom to explore different perspectives. Others are optimized for still images, making them unsuitable for video applications due to temporal inconsistencies that produce unnatural flickering effects. Additionally, some methods are computationally expensive, taking hours to process a short clip or sacrificing quality to achieve faster results. In this paper, we introduce a real-time, 3D-aware method for portrait video relighting that addresses these limitations. Our approach employs Neural Radiance Fields (NeRF) to synthesize faces with realistic lighting under novel viewpoints. We utilize a dual-encoder architecture to disentangle surface color (albedo) and shading information, allowing flexible and accurate lighting control. Furthermore, a temporal consistency network is incorporated to smooth transitions across frames, reducing flicker and enhancing visual coherence. The proposed method achieves an inference speed of 32.98 frames per second on consumer-grade hardware, outperforming existing techniques in both quality and efficiency, thus enabling high-quality, interactive 3D-aware relighting for real-world applications.



Study methodology

The proposed system introduces a novel approach to real-time 3D-aware portrait video relighting, consisting of dual-encoders and a temporal consistency network. These components work in unison to achieve high-quality, temporally stable relighting, even when changing viewpoints and lighting conditions in dynamic scenes.

- **Dual Encoders** The system employs dual encoders: an albedo encoder and a shading encoder, which separately capture surface color (albedo) and lighting (shading) information from input frames. This disentanglement allows the model to independently control the appearance and lighting conditions of the portrait, making it possible to adjust illumination without affecting the inherent facial texture. The albedo encoder retains the subject's consistent surface details, while the shading encoder adapts to various lighting effects, achieving realistic and flexible relighting. By encoding these aspects separately, the model gains fine-grained control over each factor, enhancing photorealism and adaptability under diverse lighting settings.
- **Tri-plane Representation** Inspired by Neural Radiance Fields (NeRF), the system utilizes a tri-plane representation to support high-resolution, 3D-aware relighting. The tri-plane structure encodes three key aspects of the scene: depth, albedo, and shading information, distributed across three planes. This representation not only facilitates efficient storage and computation but also enables accurate 3D modeling of faces under novel viewpoints and lighting conditions. By incorporating the tri-plane structure, the system can efficiently generate high-quality relighting results that preserve the facial geometry and texture, enhancing realism in both frontal and off-angle views.
- **Temporal Consistency Network** To address temporal stability across frames, a temporal consistency network is incorporated. This network uses cross-attention mechanisms between the albedo and shading encoders, aligning information across consecutive frames to minimize flickering and ensure smooth transitions. The cross-attention mechanism allows the network to share relevant details between frames, reducing inconsistencies and maintaining visual coherence over time. Additionally, the temporal consistency network uses synthetic training data with varying lighting and viewpoint conditions, which enhances robustness and helps the model generalize to real-world dynamic expressions. As a result, the system achieves stable and flicker-free relighting for video sequences, supporting realistic and seamless viewing experiences.

In summary, the proposed dual-encoder system with tri-plane representation and temporal consistency network enables efficient and high-quality portrait video relighting in real-time. This architecture provides users with the ability to control lighting conditions dynamically, improving visual consistency and enhancing interactivity for real-world applications.

Training Process

Dataset Preparation:

- The model is trained and tested on the INSTA dataset.
- Each frame is cropped to focus on the face, with camera poses estimated using EG3D and lighting conditions extracted using DPR.

Training Stages:

- **Stage 1:** The dual-encoder network is initially trained independently for albedo and shading extraction, with the generator kept frozen.
- **Stage 2:** After 16 million iterations, the entire model, including albedo and shading decoders along with the super-resolution module, is fine-tuned jointly to enhance consistency.

Temporal Consistency Network:

- Two views are created per individual by sampling camera poses, allowing the model to learn representations that maintain temporal stability across frames.

Optimization:

- The Adam optimizer is employed with a learning rate of 0.0001 for most parameters and 0.00005 for Transformer components.

Loss and maths involved

- **Albedo Loss** :

$$L_{\text{albedo}} = ||\hat{A} - A||1 + L\text{lpips}(\hat{A}, A) + \lambda_g||T_g - T_{\hat{g}}||_1,$$

- **Shading Loss** :

$$L_{\text{shading}} = ||\hat{S} - S||_1 + \lambda_s||T_S - \hat{T}_S||_1.$$

- **RGB Loss** :

$$L_{\text{rgb}} = ||\hat{I} - I||1 + L\text{lpips}(\hat{I}, I) + L_{\text{id}}(\hat{I}, I).$$

- **Adversial Loss** :

$$L_{\text{adv}} = -\left(E[\log D(I)] + E[\log(1 - D(\hat{I}))]\right).$$

- **Temporal Consistency Loss** :

$$L_{\text{short}} = M_s \sum_{\omega \in \{I,A,S\}} L_{\text{lpips}}(\omega^i - \tilde{\omega}^{i-1}),$$

Results

Performance:

- Runs at 32.98 fps on consumer-grade GPUs, offering realtime relighting for portrait videos.

Quality Comparison:

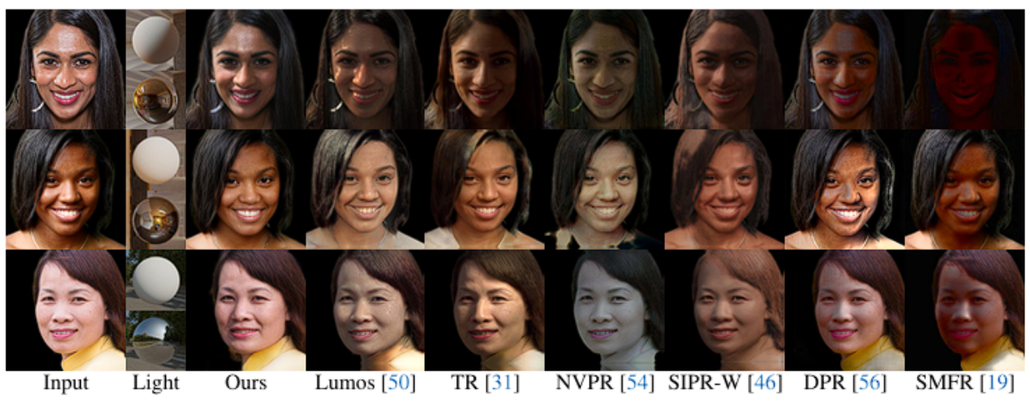
- Outperforms existing methods (e.g., DPR, SMFR) in the following metrics:
  - Lighting accuracy
  - Identity preservation
  - Warping error
  - Temporal consistency

Quantitative Metrics:

- Lighting Error: 0.771 (improved over baselines)
- Lighting Instability: 0.253 (reduced flickering)
- Identity Preservation: 0.5396 (high fidelity to original features)

Qualitative Outcomes:

- Relighting results show natural lighting transitions and reduced artifacts even in complex lighting conditions (e.g., side lighting).



Conclusions

- We introduced a real-time 3D-aware method for portrait video relighting and novel view synthesis. Our method can recover coherent and consistent geometry and relight the video under novel lighting conditions for a given facial video. This method combines the benefits of a re lightable generative model, i.e., disentanglement and con trollability, to capture the intrinsic geometry and appearance of the face in a video and generate realistic and consistent videos under novel lighting conditions.
- **Limitations** One of the limitations of our method is that it fails to model glares on the eyeglasses, as shown in the rightmost column of Figure 4. Future enhancements could benefit from incorporating advanced reflection and refrac tion modeling techniques

omshri21@iitk.ac.in