Real-time 3D-aware Portrait Video Relighting

Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, Lin Gao

Motivation

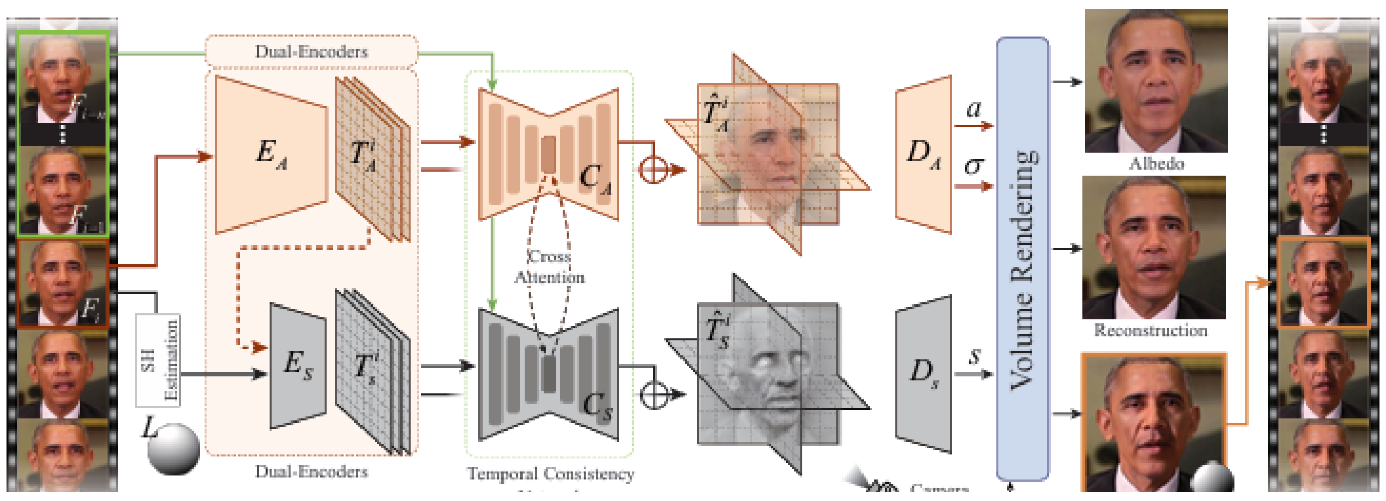
Drawbacks of the current relighting techniques:

- Limited to **original viewpoints**, restricting creative & practical applications
- Issues with temporal consistency, causing **flickering** in video sequences
- Very high **computational costs**, preventing real-time applications

Methodology

The proposed system consists of dual-encoders and a temporal consistency network:

- **Dual Encoders**: Use albedo and shading encoders to capture surface color (albedo) and lighting (shading) information. This disentanglement allows independent control over appearance and lighting.
- **Tri-plane Representation**: Inspired by Neural Radiance Fields (NeRF), the tri-plane structure supports high-resolution 3D-aware relighting by encoding scene depth, albedo, and shading information.
- **Temporal Consistency Network**: Utilizes cross-attention between albedo and shading encoders, enhancing temporal stability across frames and reducing flicker.



Training process

Dataset Preparation:

The model is trained and tested on the **INSTA dataset**, consisting of 31,079 portrait frames. Each frame is cropped to focus on the face, with camera poses estimated using **EG3D [5]** and lighting conditions extracted using DPR.

Training Stages:

- **Stage 1:** The dual-encoder network is trained independently for albedo and shading extraction, with the generator frozen.
- **Stage 2:** After 16 million iterations, the entire model (albedo and shading decoders plus the super-resolution module) is jointly fine-tuned for improved consistency.
- **Temporal Consistency Network:** Two views are created per individual by sampling camera poses, allowing the model to learn temporally stable representations.

Optimization:

The Adam optimizer is used, with a learning rate of **0.0001** for most parameters and 0.00005 for Transformers.

Training Duration:

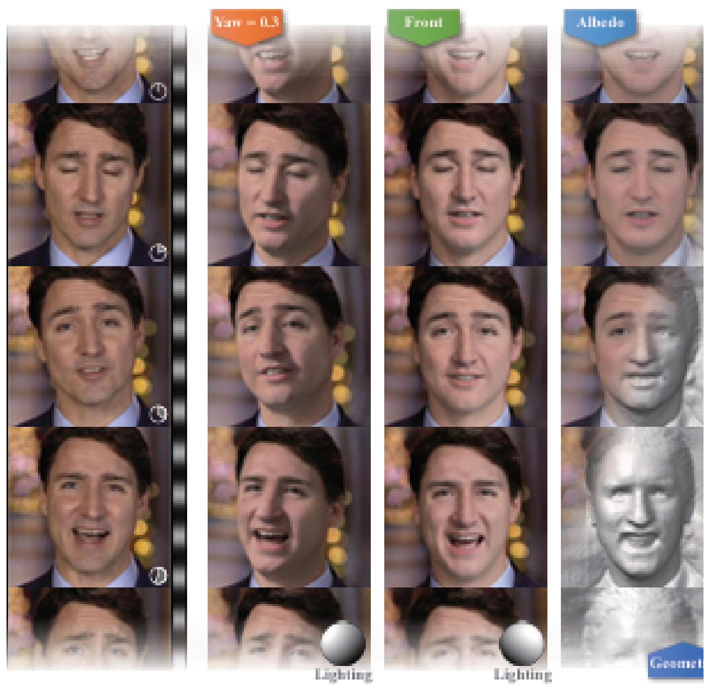
Training on 8 NVIDIA Tesla V100 GPUs with a batch size of 32 takes approximately **30 days** to converge.

Conclusion

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.

Objective

- Develop a **3D-aware, real-time** method for relighting portrait videos.
- Enable **viewpoint changes** and **realistic lighting** adjustments for portrait videos with consistent quality.
- Achieve temporal stability and **reduce flickering** in relit videos, making it ideal for dynamic applications in AR/VR.



Key Equations Used

Albedo Loss

$$\mathcal{L}_{\text{albedo}} = ||\hat{A} - A||_1 + ||\hat{A}_r - A_r||_1 + \mathcal{L}_{\text{lpips}}(\hat{A}, A) + \mathcal{L}_{\text{lpips}}(\hat{A}_r, A_r) + \lambda_g ||\hat{T}_g - T_g||_1,$$

Shading Loss

$$\mathcal{L}_{\text{shading}} = ||\hat{S} - S||_1 + \lambda_s ||\hat{T}_S - T_S||_1,$$

RGB Loss

$$\mathcal{L}_{\text{rgb}} = ||\hat{I} - I||_1 + ||\hat{I}_r - I_r||_1 + \mathcal{L}_{\text{lpips}}(\hat{I}, I) + \mathcal{L}_{\text{lpips}}(\hat{I}_r, I_r) + \lambda_f ||\hat{I}_f - I_f||_1 + \mathcal{L}_{id}(\hat{I}, I),$$

Adversarial Loss

$$\mathcal{L}_{\text{adv}} = -(\mathbb{E}[\log D(I)] + \mathbb{E}[\log D(I_r)] + \mathbb{E}[\log(1 - D(\hat{I}))] + \mathbb{E}[\log(1 - D(\hat{I}_r))]).$$

Total Loss for training Dual- Encoders

$$\mathcal{L} = \lambda_{\text{albedo}}\mathcal{L}_{\text{albedo}} + \lambda_{\text{shading}}\mathcal{L}_{\text{shading}} + \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}},$$

Temporal Consistency Loss

$$\mathcal{L}_{\text{short}} = M_s^i \sum_{\omega \in \{\hat{I}, \hat{I}_r, \hat{A}, \hat{A}_r, \hat{S}\}} \mathcal{L}_{\text{lpips}}(\omega^i - \widetilde{\omega}^{i-1}),$$

Results

- Performance:
  - Runs at **32.98 fps** on consumer-grade GPUs, offering real-time relighting for portrait videos.
- Quality Comparison:
  - **Outperforms** existing methods (e.g., **DPR, SMFR**) in metrics of lighting accuracy, identity preservation, warping error, and 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).