Implementation Details & Dataset Information

Name: Deepanshu

Roll No.: 210311

1 Introduction

This project is to make a model with two parts called dual encoders that can invert videos and change lights in them. The main idea is to split features from an image into two different parts called "tri-planes" – one for texture (albedo) and one for lighting (shading). This way, the model tries to stay stable for each video frame, and transformers help reduce flickering.

2 Implementation Details

2.1 Model Design

- **Dual Encoders:**
  - **Albedo Encoder:** It's a Vision Transformer (ViT) [12] that takes an RGB image with coordinate info to predict the albedo tri-plane, capturing shape and texture.
  - **Shading Encoder:** Uses CNNs plus some StyleGAN [22] layers to create a shading tri-plane based on albedo tri-plane and lighting conditions (spherical harmonics).
- **Temporal Consistency Network:**
  - To prevent flickering, we use transformers with cross-attention, which lets albedo and shading parts talk to each other to stay smooth between frames.

2.2 Training Process

The model is trained using multiple loss functions:

- **Albedo Loss:** Makes predicted albedo image look like the real one.
- **Shading Loss:** Compares predicted and real shading components.
- **RGB Loss:** Keeps predicted RGB images close to real images with perceptual and identity loss.

- **Adversarial Loss:** Uses a discriminator to make the output look more realistic.
- **Temporal Consistency Loss:** Makes frames smooth by checking flow between frames.

Training is done in three stages:

- (1) Train albedo encoder.
- (2) Train both encoders separately.
- (3) Train everything together.

3 Dataset Information

- **Synthetic Data:** Synthetic data is created with random camera and light angles, similar to real videos. Flickering effects are added for de-flickering.
- **Data Augmentation:** Data is augmented with view and noise changes, so the model gets used to video transitions.
- **Testing Data:** Testing uses high-res face videos with different light and camera angles, with labels to compare outputs.