High-fidelity Face Tracking for AR/VR via Deep Lighting Adaptation

Lele Chen[1,2] Chen Cao[1] Fernando De la Torre[1] Jason Saragih[1] Chenliang Xu[2] Yaser Sheikh[1]
[1] Facebook Reality Labs [2] Univeristy of Rochester

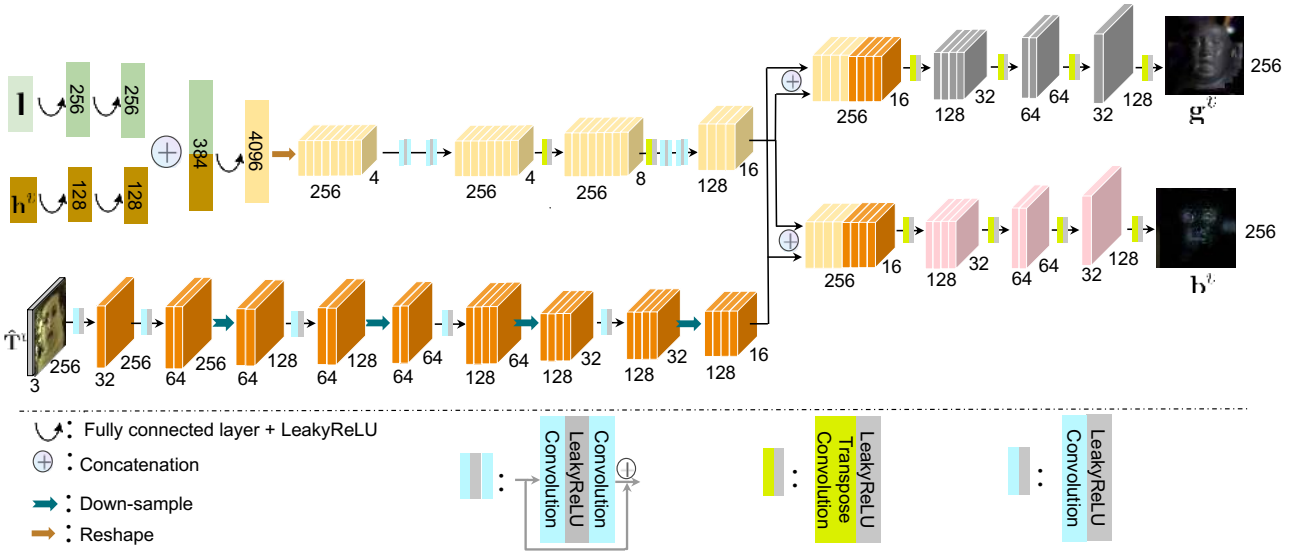

Figure 1. The detailed network structure of our lighting transfer network (G).

In this supplementary file, we explain the network details of our lighting model.

A. Network Structure

We present the detailed network structure in Fig. 1.

B. Inputs and Outputs

The lighting code $\mathbf{l}$ is a pre-defined vector when we train G on light-stage data, and is a learnable vector when we refine G on in-the-wild video frames. During training on light-stage data, the lighting direction is encoded by the position of the non-zero element in $\mathbf{l}$, and the lighting color is encoded by the value of the non-zero element in $\mathbf{l}$. The view-dependent head pose $\mathbf{h}^v \in \mathbb{R}^6 = \{\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z, \mathbf{v}_x^v, \mathbf{v}_y^v, \mathbf{v}_z^v\}$, where $\mathbf{r} = \{\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z\}$ and $\mathbf{v}^v = \{\mathbf{v}_x^v, \mathbf{v}_y^v, \mathbf{v}_z^v\}$ are rigid head rotation and viewpoint vector, respectively. The fully-lit texture $\hat{\mathbf{T}}^v$ is obtained from DAM decoder, and we down-sample it to the size of $3 \times 256 \times 256$.

The outputs are the gain and bias map $\mathbf{g}^v, \mathbf{b}^v$, and we upsample the output $\mathbf{g}^v, \mathbf{b}^v$ back to the size of $3 \times 1024 \times 1024$ by bilinear interpolation.