

Temporally Consistent Relighting for Portrait Videos

Sreenithy Chandran¹, Yannick Hold-Geoffroy², Kalyan Sunkavalli², Zhixin Shu², and Suren Jayasuriya¹

¹Arizona State University

²Adobe Research

In this supplementary document, we outline the training procedure and the architecture of the network. We refer the reader to our accompanying video for various relighting results.

Neural Network Architecture

In this section, we detail our network architecture, which comprises four major steps: warp, relight, unwarp, and consistency. Let us denote the batch size as B , and the height and width of the frames to be H, W . The Table 1 details the tensor sizes at the various stage of the architecture.

Step	Input	Output
Input	$[B, 3, H, W]$	
Warp	$[B, 3, H, W]$	$[B, 3, 512, 512]$
RGB to LAB	$[B, 1, 512, 512]$	$[B, 3, 512, 512]$
Relighting	$[B, 1, 512, 512]$	$[B, 1, 512, 512]$
LAB to RGB	$[B, 1, 512, 512]$	$[B, 3, 512, 512]$
Unwarp	$[B, 3, 512, 512]$	$[B, 3, H, W]$
Blend	$[B, 3, H, W]$	$[B, 3, H, W]$
Consistency	$[B, 12, H, W]$	$[B, 3, H, W]$

Table 1. Details about the various steps in the network

Table 2, shows the blind temporal consistency architecture [1]. The layers $c1a, c1b, c2a, c2b, c3, s3$ are convolutional layers, $R1, R2, R3, R4, R5$ are residual blocks, $LSTM$ is the convolutional LSTM layer [2] with hidden size 128, $s1, s2$ are upsampling convolutional layers. Every convolutional layer is followed by a ReLU activation. The P_t and P_{t-1} which are the relit frames at time t and $t - 1$ are concatenated into one input and fed to $c1b$, and O_{t-1} is the previously stabilized relit frame at time $t - 1$ output by the network at a previous iteration and P_t are concatenated and fed as input to $c1a$. Output of $c1a, c1b$ is fed as input to $c2a, c2b$ respectively. Output of $c2a, c2b$ are concatenated channel-wise and fed to the residual blocks. Output of $c2a$ is concatenated with $s1$ and fed to $s2$. Output of $c1a$ is concatenated with $s2$ and fed to $s3$.

layer name	in ch	out ch	k-sz
c1a	6	32	7
c1b	6	32	7
c2a	32	64	3
c2b	32	64	3
c3	128	128	3
R1	128	128	-
R2	128	128	-
R3	128	128	-
R4	128	128	-
R5	128	128	-
LSTM	128	128	-
s1	128	64	3
s2	128	32	3
s3	64	3	7
tanh	3	3	-

Table 2. Details about the consistency architecture showing the number of input and output channels and kernel size.

References

- [1] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018.
- [2] Xingjian Shi, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214*, 2015.