Supplementary Materials

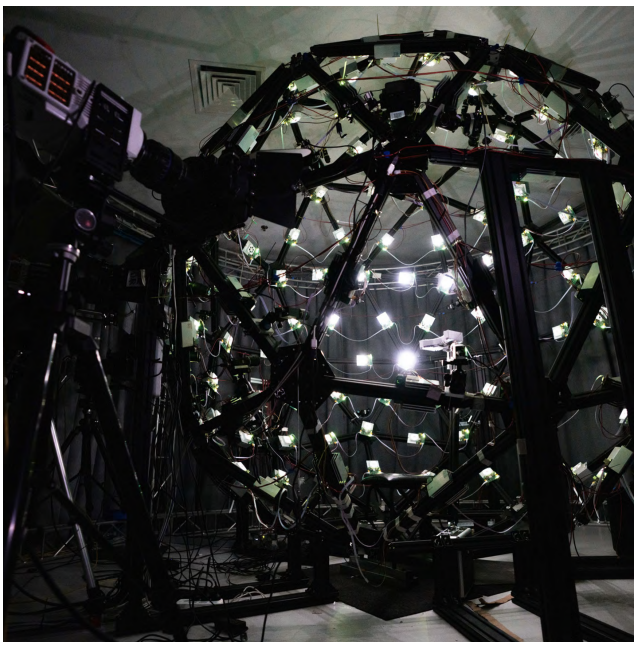

Figure 1. The demonstration of our capturing system. Cameras and lights are arranged on a spherical structure.

Figure 2. The samples of our captured data. (a) to (e) are samples of OLAT images; (f) is a full-lit frame image for calculating the optical-flow. The bottom-left illustrates the corresponding lighting conditions on our system.

1. Hardware architecture and capture settings

To recover the 4D reflectance fields of dynamic portrait, we build up a light stage device with 96 LED light sources and a stationary 4K ultra-high-speed camera at 1000 fps, resulting in a temporal OLAT image set at 25 fps, so as to provide fine-grained facial perception.

Under such a dynamic capture setting, the target performer can freely speak, translate and rotate in a certain range to provide a continuous and dynamic OLAT image sets sequence. However, one of the most challenging issues is that the motion of the captured target along with data acquisition will cause misalignment, leading to blurriness in the OLAT image sets, making the data post-processing more challenging. We conquer such limitations using an optical flow algorithm and further retrieve results at a higher frame rate using densely-acquired homogeneous full-lit frames.

Our hardware architecture is demonstrated in Fig. 1, which is a spherical dome of a radius of 1.3 meters with 96 fully programmable LEDs and a 4K ultra-high-speed camera. The LEDs are independently controlled via a synchronization control system and evenly localized on the doom to provide smooth lighting conditions.

Single PCC (Phantom Capture Camera) is leveraged with a global shutter setting, generating roughly 6 TB of data in total. For each data session, we collect video sequences in 25 seconds rather than isolated frames with 700 us exposure time and 1000 EI, which allows a lower noise floor. In practice, we can simultaneously control the PCC along with the LEDs system.

During capture, all the 96 LEDs follow the patterns shown in Fig. 2.

During the capture period, the high-speed camera synchronizes with the 96 LEDs at 1000 fps and outputs an 8-bit Bayer pattern color image stream at a resolution of 2048×1440.

Note that it is required at least 0.1s to acquire a complete dynamic session with 96 LEDs. Such duration causes misalignment, making it challenging to handle blurriness in the dataset and the low frame rate gives rise to inconsistent dy-

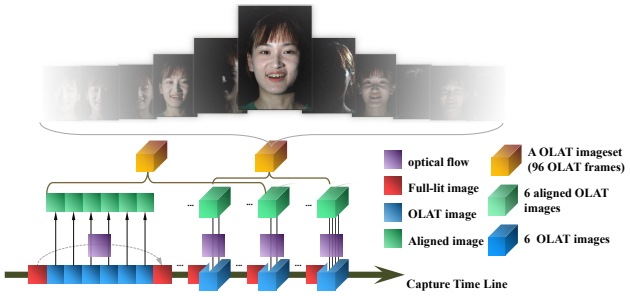

Figure 3. The illustration of capture timing and frame alignment. our method computes optical flows between full-lit images and warps OLAT images accordingly. The "overlapping" strategy allow us to reuse a same OLAT image in different OLAT image set so that we can achieve higher capture frame rate.

namic facial capture results. Inspired by the approach [1], we interleave "tracking frames" into the capture sequence during every six images rather than a complete session to conquer such drawbacks and cast the tracking frames as references, as shown in Fig. 3.

Instead of capturing an image with homogeneous illumination for every 96 images, we capture an image for tracking purposes every 6 images. This capture strategy allows us to align the OLAT data between 14 consecutive groups of full-bright frames in any pose to the middle frame with optical flow. It is equivalent to that the image between every two homogeneous illumination frames is multiplexed thirteen times to enhance the final optical flow result.

2. Normalization Operation Process

Note that when the source and target actors are different, their facial geometry will be different even in similar facial expressions, which will manifest in facial characteristics, such as eye size, nose length, mouth curvature, etc. We expect the distribution of conditioning feature maps generated from the source actor is similar to one of the target actor, so as to preserve target-aware appearance rendering results.

Specifically, we assume parameters as random variables which follow the normal distributions. For each parameter, we normalize both the mean and variance of the source actor's distribution $\mathbf{X}_s \sim N(\mu_s, \sigma_s^2)$ to be the same as the target actor's distribution $\mathbf{X}_t \sim N(\mu_t, \sigma_t^2)$. Thus, the normalized parameter $\hat{X}_s$ can be formulated as:

$$\hat{X}_s = (\sigma_t \oslash \sigma_s) \circ (\mathbf{X}_s - \mu_s) + \mu_t \sim N(\mu_t, \sigma_t^2), \quad (1)$$

where $\oslash$ and $\circ$ are element-wise division and product respectively. In our implementation, we regard the head pose θ and facial expression ϕ of FLAME parameters as two individual random variables and use the Eqn. 1 for normalization.

Thus, the conditioning feature maps $\mathbf{F}_t$ at time t are formulated as:

$$\mathbf{F}_t = \{\tilde{\mathbf{I}}_t^d, \tilde{\mathbf{I}}_t^c, \tilde{\mathbf{I}}_t^l\}, \quad (2)$$

where $\tilde{\mathbf{I}}_t^d, \tilde{\mathbf{I}}_t^c = \Pi(G(\beta, \tilde{\theta}, \tilde{\phi}), \mathbf{H})$ and Π is the rasterization and texture mapping function; $\tilde{\theta}$ and $\tilde{\phi}$ are normalized parameters.

3. Network Architecture

We adopt the U-Net architecture to the proposed translation network as illustrated in Fig. 4. Our network inferences both the facial reflectance field, the normal, and the albedo simultaneously to facilitate portrait video composition applications. The proposed network consists of an encoder and a decoder module. The encoder extracts multi-scale latent representations of conditioning feature maps, while the decoder module generates the reflectance field, the normal, and the albedo.

Such a multi-task framework enforces the network to learn contextual information of the target actor, so as to produce more detailed reflectance fields.

4. Quantitative comparisons

For quantitative comparison, we evaluate our rendering results via three various metrics: signal-to-noise ratio (**PSNR**), structural similarity index(**SSIM**) and mean absolute error (**MAE**) to compare with existing state-of-art approaches. For comparison, we generate a sequence of relighted portraits from all reference views which are evenly spaced in range of illumination angles. As shown in Fig. 5 due to the high-quality reflectance field inference and fully disentanglement, our method enables entire photo-realistic video portraits synthesis under various illumination conditions, making the rendering quality of any temporal sequence surpassing other baselines with higher PSNR, SSIM and lower MAE.

5. Application: Virtual Conference

By utilizing our generated dynamic facial reflectance field, we can achieve a relightable virtual conference as shown in Fig. 6. No matter whether the user is dressing casually with messy hair, or lying on the sofa with a cup of coffee, the user can appear decently in front of others in suits. The conference background can be changed arbitrarily, the most impressive effect is that the light conditions on the generated user in suits will perfectly match the environment. Furthermore, if the user is walking on the street in a rush with a shaking camera, or even he is not looking and the camera, by explicitly controlling the pose parameters, our approach can always let the user look like attending the conference naturally.

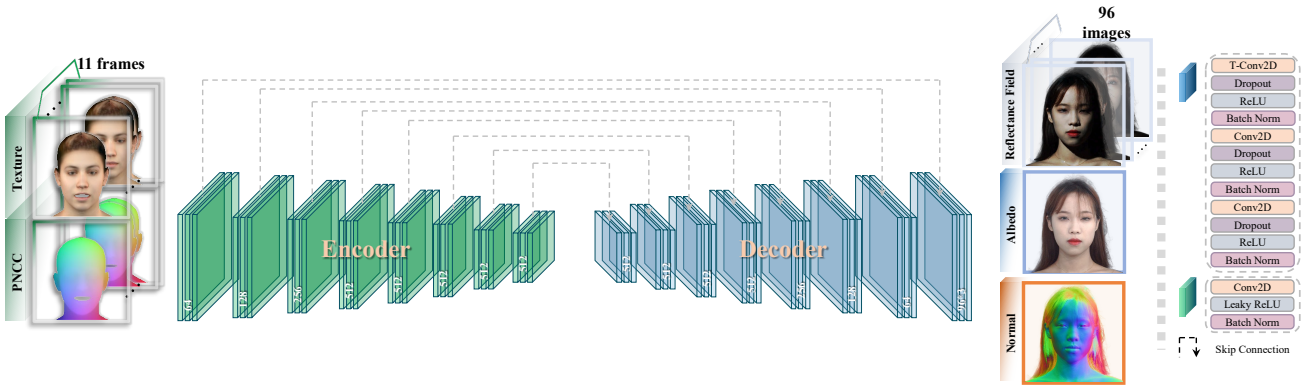

Figure 4. Our multi-frame multi-task architecture design for our rendering-to-video network.

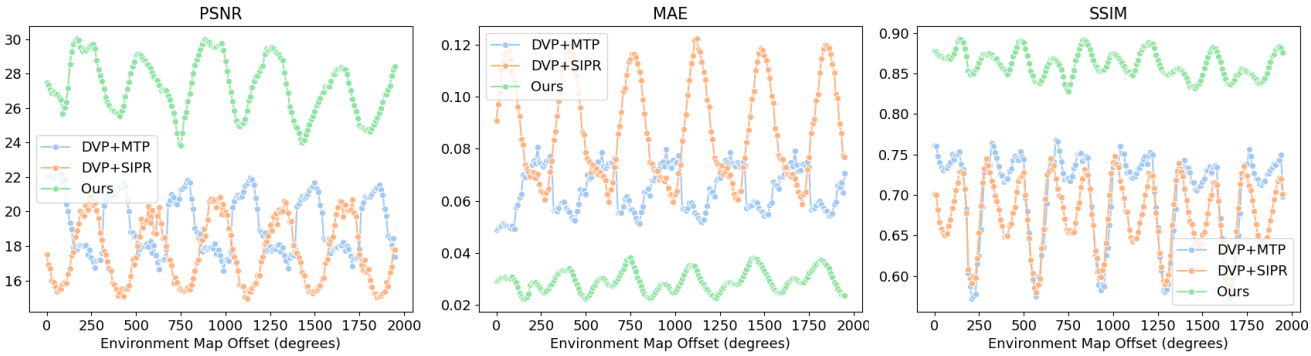

Figure 5. Quantitative comparisons on the test reference data with ground truth. Our results are more realistic and closer to ground truth.

References

[1] Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–21, 2020. 2

Figure 6. We demonstrated the application of our algorithm in the virtual conference. Users can participate in video conferences on formal occasions in a widerrange of scenarios. Through the control of pose, we can realize that the user still presents a decent participation state in scenes such as walking and notlooking at the camera.