Supplementary Material
RelightableHands: Efficient Neural Relighting of Articulated Hand Models

Shun Iwase[1,2*] Shunsuke Saito[2] Tomas Simon[2]
Stephen Lombardi[2] Timur Bagautdinov[2] Rohan Joshi[2]
Fabian Prada[2] Takaaki Shiratori[2] Yaser Sheikh[2] Jason Saragih[2]
[1]Carnegie Mellon University [2]Reality Labs Research

1. Network Architecture

In this section, we provide the network architecture details and hyperparameters for the teacher and student models. In addition, we explain the modifications we made to DRAM [1] for a fair comparison to our proposed method.

1.1. Teacher Model

The joint feature encoder $\mathcal{J}_t\left(\theta\right)$ of the teacher model first repeats and tiles the hand pose $\theta \in \mathbb{R}^{25}$ across the UV space and $\theta' \in \mathbb{R}^{25\times64\times64}$ is obtained Given the hand pose feature θ', a 2-layer convolutional neural network (CNN) with channel sizes $(16, 64)$ outputs a joint feature $\mathcal{J}_t\left(\theta\right) \in \mathbb{R}^{64\times64\times64}$. $\mathcal{A}_{\text{OLAT}}$ adopts a U-Net [4] architecture which takes $wS \times wS \times 7S$ per-light feature as an input. The U-Net encoder is a 4-layer CNN with channel sizes $(7S, 64, 64, 64)$. Here, the joint feature is concatenated with the output feature from the U-Net encoder and passed to the U-Net decoder. The U-Net decoder is a 4-layer CNN with a skip connection from the U-Net encoder, with channel sizes $(128, 128, 128, 4S)$. We use bilinear interpolation to downsample and upsample the features at each layer. The first $3S$ output channels encode raw RGB volumetric texture $\mathbf{T}_k^i$, and the last S channels the shadow map $\mathbf{S}_k^i$. The OLAT texture $\mathbf{C}_k^i$ for the i-th light is computed as follows;

$$\mathbf{C}_k^i = \sigma(\mathbf{S}_k^i)\left(\text{ReLU}\left(\lambda_s\mathbf{T}_k^i\right) + \lambda_b\right) \in \mathbb{R}^{S\times S\times 3S}, \quad (1)$$

where $\sigma(\cdot)$ is a sigmoid function, $\text{ReLU}(x) = \max(0, x)$, λ_s is a scale parameter, and λ_b is a bias parameter. In our experiments, we set λ_s to 25 and λ_b to 100.

1.2. Student Model

The joint feature encoder $\mathcal{J}_s\left(\theta\right)$ of the student model is the same architecture as $\mathcal{J}_t\left(\theta\right)$. However, the student model only has a texture decoder network $\mathcal{A}_{\text{env}}$, which takes a $76 (= 3 + 9 + 64) \times 128 \times 128$ feature as an input. The texture decoder is an 8-layer CNN with channel sizes $(76, 256, 256, 128, 128, 64, 64, 32, 3S)$, outputting raw RGB volumetric texture. Given the output raw RGB texture $\mathbf{T}_k$, the relit texture $\mathbf{C}_k$ is expressed as

$$\mathbf{C}_k = \text{ReLU}\left(\lambda_s\mathbf{T}_k\right) + \lambda_b \in \mathbb{R}^{S\times S\times 3S}, \quad (2)$$

where the same hyperparameters λ_s and λ_b are the same as for the teacher model.

1.3. DRAM [1]

Unlike DRAM [1], our hand model is animated by the hand pose parameter instead of facial expressions. Therefore, we replace the latent code z in their hyper-network architecture with the hand pose parameter θ. In addition, we rotate the environment map to align with the coordinate system of the hand's root joint before it is given to the hyper-network.

2. Comparison to GGX Microfacet BRDF [5]

Our teacher model directly estimates an OLAT rendering from localized light and view directions and shadow maps. An alternative approach is to adopt a physically-based BRDF model and render with the rendering equation. To validate the effectiveness of our neural rendering approach, we compare the teacher model against a physically-based rendering baseline that is heavily inspired by the recent success of Relighting4D [2]. This physically-based baseline predicts material properties such as albedo, roughness, and a localized normal map per voxel and feeds these into GGX microfacet BRDF [5] to obtain the RGB textures.

We adopt the same network architecture as the original teacher model with the joint feature encoder for deep shadow map and albedo estimation except for the first and output channel sizes. Concretely, we give the local light directions $\mathbf{F}_v^k$ and visibility map $\mathbf{V}(l_i)_k$ as an input to the shadow map U-Net to obtain the per-light shadow map $\mathbf{S}_k^i \in \mathbb{R}^{S\times S\times S}$. Also, we feed the average texture $\bar{\mathbf{T}}$ into the albedo U-Net to acquire the albedo $\mathbf{A}_k \in \mathbb{R}^{S\times S\times 3S}$. Since MVP [3] does not provide normal information, we learn normal and roughness decoders with the channel sizes

	Subject 1						Subject 2					
	MSE ($\times10^{-3}$) ↓			SSIM ↑			MSE ($\times10^{-3}$) ↓			SSIM ↑		
	Right	Left	Both	Right	Left	Both	Right	Left	Both	Right	Left	Both
GGX [2,5]	15.2151	19.5409	59.7985	0.9591	0.9579	0.9201	22.3159	30.0019	65.5644	0.9145	0.9231	0.8695
Ours	**4.9126**	**5.8608**	**15.7589**	**0.9790**	**0.9805**	**0.9536**	**8.8205**	**7.9357**	**22.3559**	**0.9541**	**0.9559**	**0.9075**

Table 1. **Quantitative comparison of the teacher model.** We measure the MSE and SSIM metrics on the right, left, and two-hand sequences. The result shows that our method significantly outperforms the physically-based rendering baseline based on Relighting4D [2].

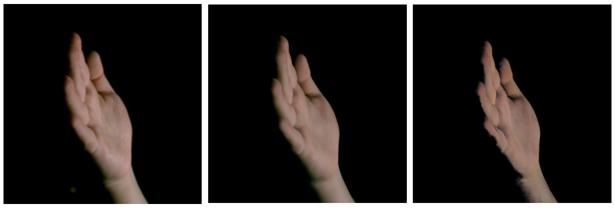

Figure 1. **Comparison with GGX microfacet BRDF model.** While the quality of relighting with physically-based relighting is bounded by the underlying BRDF model, our neural relighting model learns to synthesize photorealistic light transport effects without specifying material parameters.

$(128, 128, 128, 4S)$ and the same architecture as the teacher model. GGX microfacet BRDF [5] takes the estimated albedo $\mathbf{A}_k$, roughness $\mathbf{R}_k \in \mathbb{R}^{S\times S\times S}$, and normal $\mathbf{N}_k \in \mathbb{R}^{S\times S\times 3S}$ as input to compute RGB texture $\mathbf{T}_k$. Using the shadow map and the raw RGB texture, the predicted texture of the k-th primitive is expressed simply by

$$\mathbf{C}_k^i = \sigma(\mathbf{S}_k^i) f\left(\mathbf{A}_k, \mathbf{R}_k, \mathbf{N}_k\right) = \sigma(\mathbf{S}_k^i)\mathbf{T}_k \in \mathbb{R}^{S\times S\times 3S}, \tag{3}$$

where $f\left(\mathbf{A}_k, \mathbf{R}_k, \mathbf{N}_k\right)$ is the element-wise GGX microfacet BRDF function. To train the network, the same hyperparameters and loss functions are used.

Table 1 shows that our neural relighting outperforms the physically-based baseline by a large margin. This experimental result demonstrates that the quality of physically-based rendering is bounded by the underlying parameterization, and the GGX microfacet BRDF is unable to represent complex transmissive effects such as subsurface scattering. In contrast, the proposed neural relighting approach successfully models global light transport effects. We show the qualitative comparison between our method and the physically-based baseline in Figure 1.

3. Runtime Analysis

The student model runtime for a single hand mainly comprises 0.3 ms for joint feature decoding, 2.7 ms for ray tracing, 1.5 ms for grid sampling from 3D to the UV space, and 15.8 ms for texture decoding on NVIDIA V100.

4. Qualitative Results

4.1. Teacher Model

Figures 2 to 7 show the results of teacher model with directional lighting, near-field lighting, and environment map rendering for Subject 1 and 2. Note that the hand poses and lighting conditions are randomly sampled from the validation subset. These qualitative results clearly demonstrate the high-fidelity relighting of hand models with various poses and illuminations. Please refer to the supplemental video for animation results.

4.2. Student Model

Figures 8 and 9 exhibit the high-fidelity renderings generated by the student model in real-time. Our visibility-aware diffuse and specular features enable the generalization to unseen environment maps and hand poses. Moreover, our method is able to reproduce faithful details such as specularity and soft shadows.

References

[1] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *TOG*, 2021. 1

[2] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *ECCV*, 2022. 1, 2

[3] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *TOG*, 2021. 1

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *MICCAI*, 2015. 1

[5] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. Microfacet models for refraction through rough surfaces. In *ESGR*, 2007. 1, 2

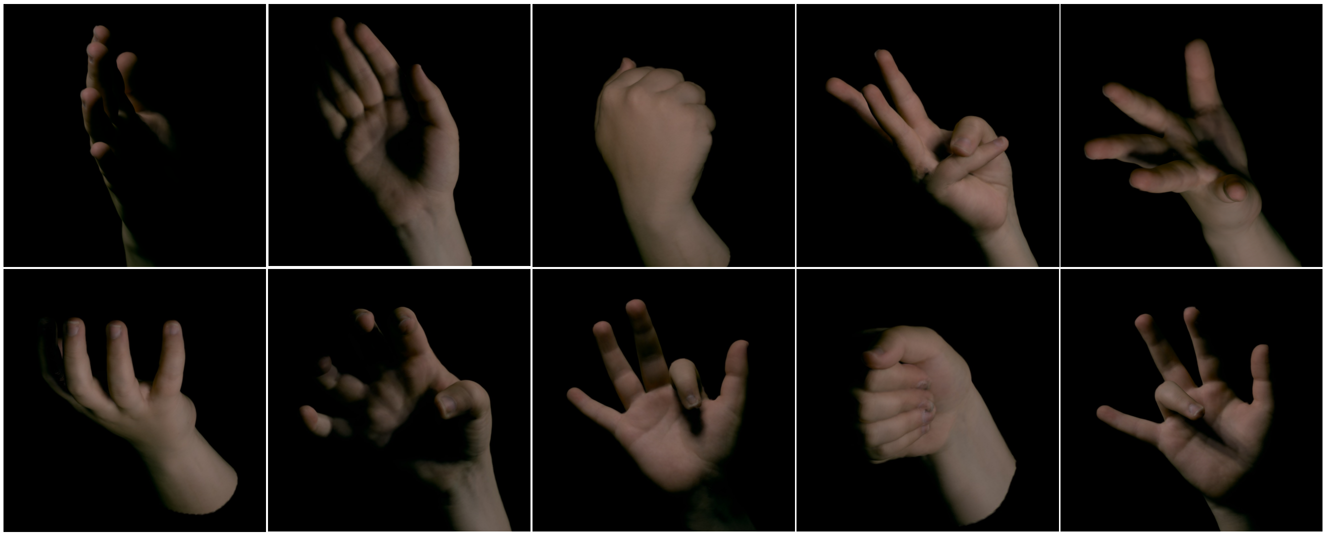

Figure 2. **Qualitative results of the teacher model with directional lighting.** Each image is generated by a randomly sampled pose and lighting condition from the test dataset of Subject 1.

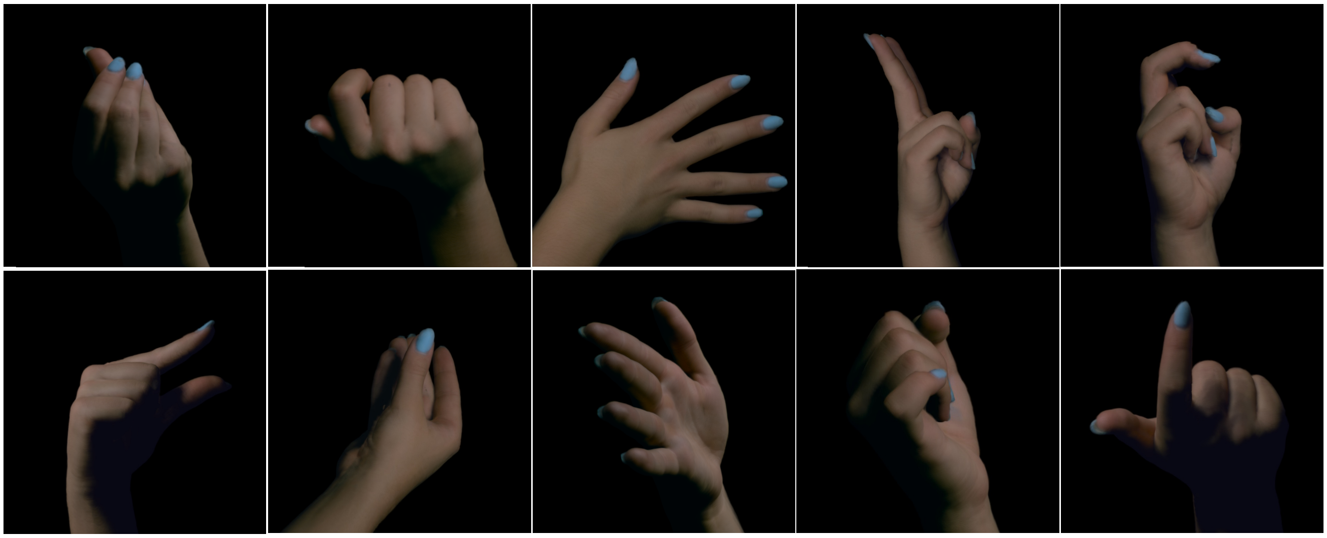

Figure 3. **Qualitative results of the teacher model with directional lighting.** Each image is generated by a randomly sampled pose and lighting condition from the test dataset of Subject 2.

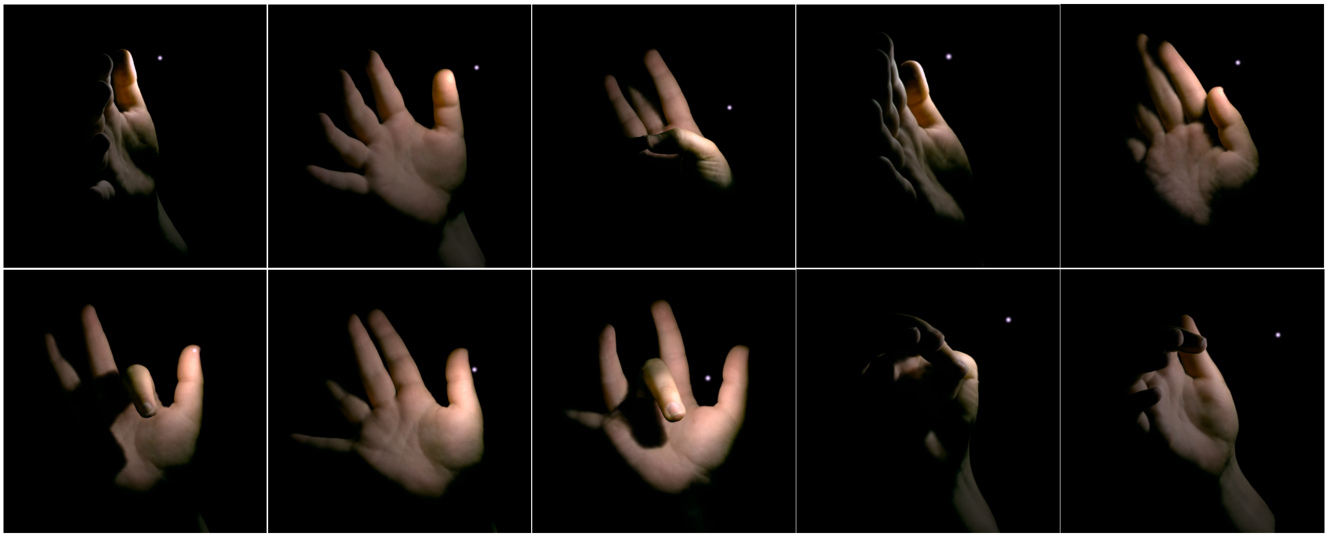

Figure 4. **Qualitative results of the teacher model with near-field lighting.** We sample the poses from the validation dataset of Subject 1. Note that we set the light locations around the fingers manually and the white dot represents the projected light location in the 2D image space.

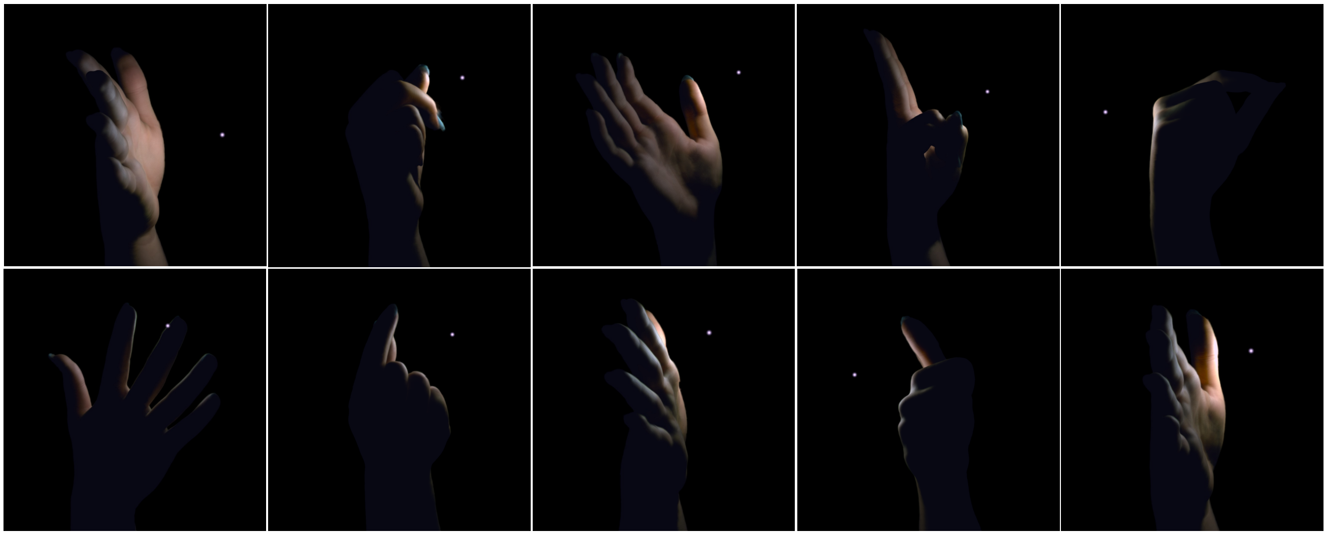

Figure 5. **Qualitative results of the teacher model with near-field lighting.** We sample the poses from the validation dataset of Subject 2. Note that we set the light locations around the fingers manually and the white dot represents the projected light location in the 2D image space.

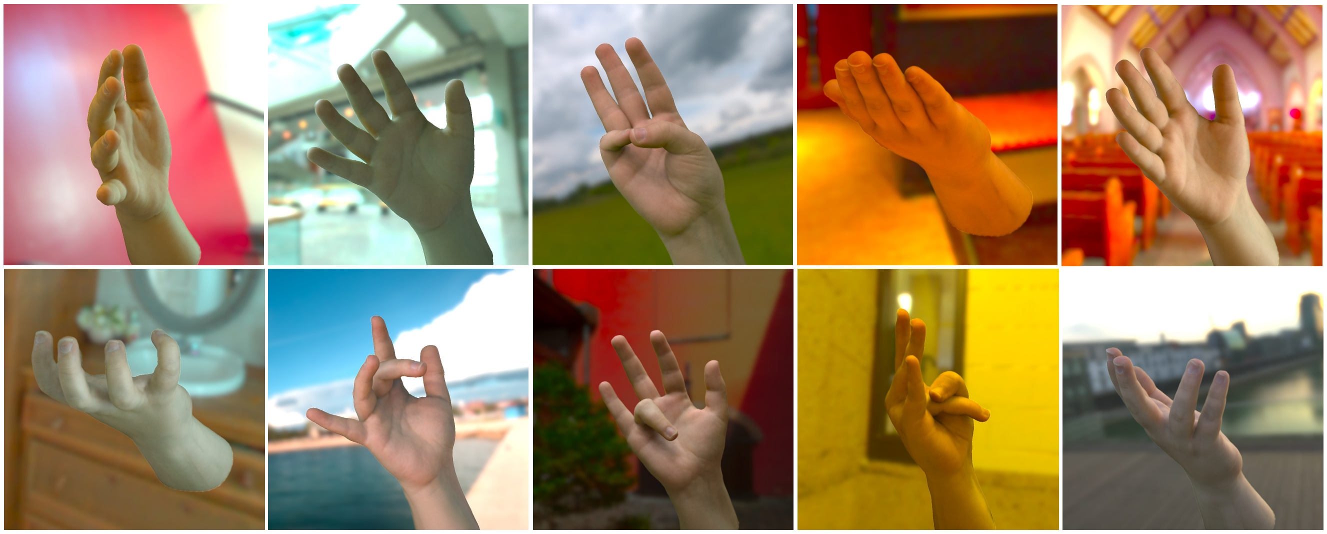

Figure 6. **Qualitative results of the teacher model with environment map lighting.** We sample the poses from the validation dataset of Subject 1.

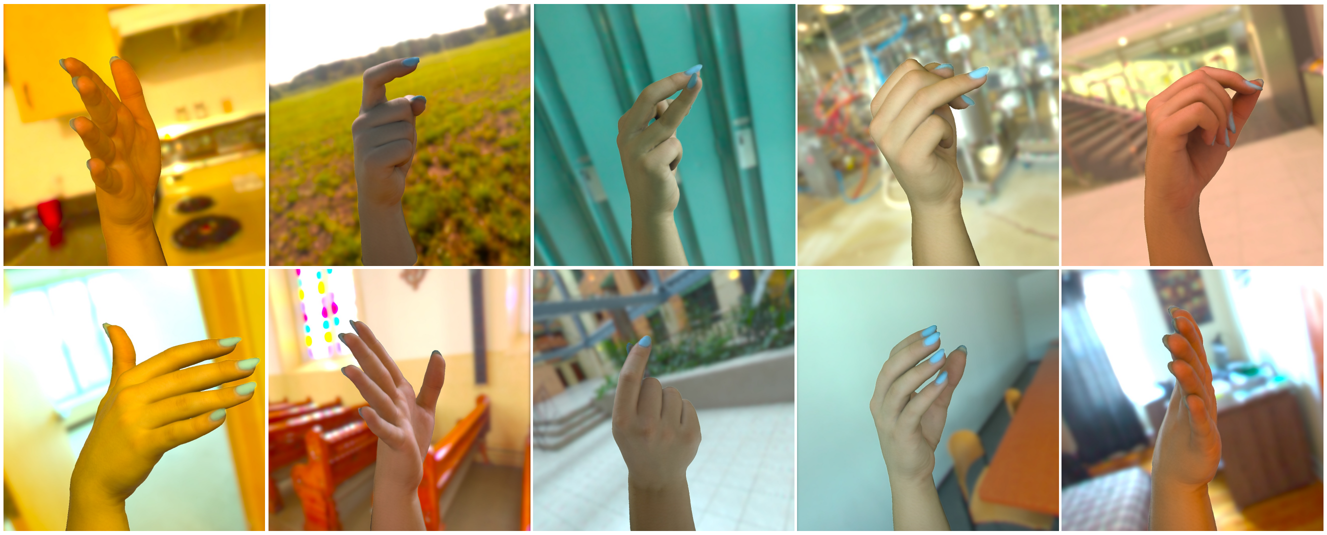

Figure 7. **Qualitative results of the teacher model with environment map lighting.** We sample the poses from the validation dataset of Subject 2.

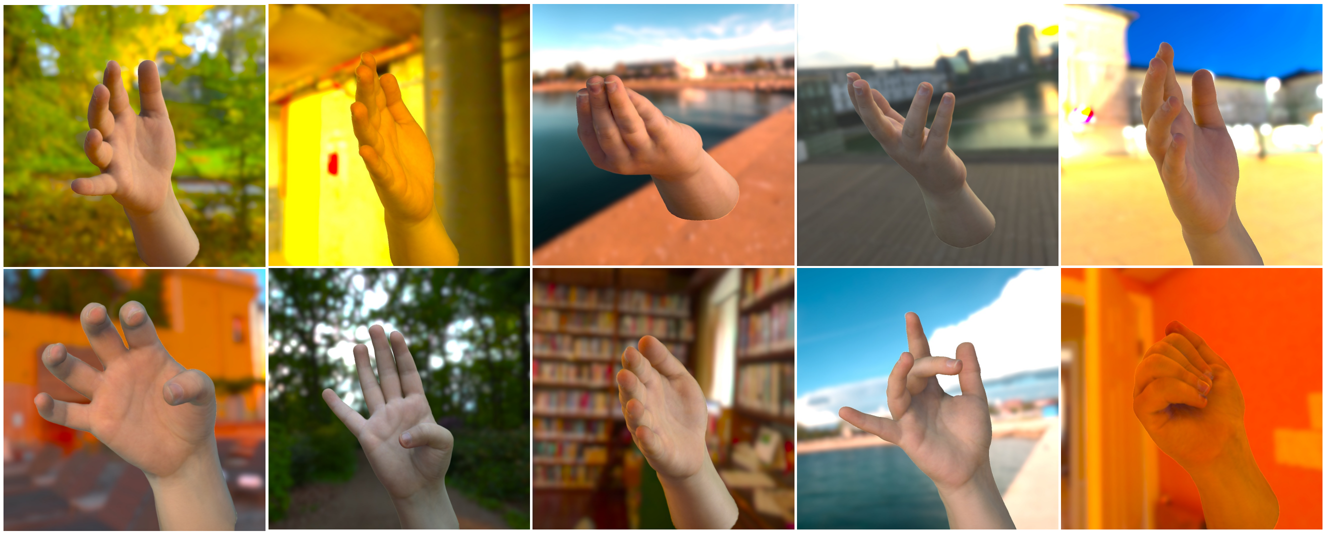

Figure 8. **Qualitative results of the student model with environment map lighting.** We sample the poses from the validation dataset of Subject 1.

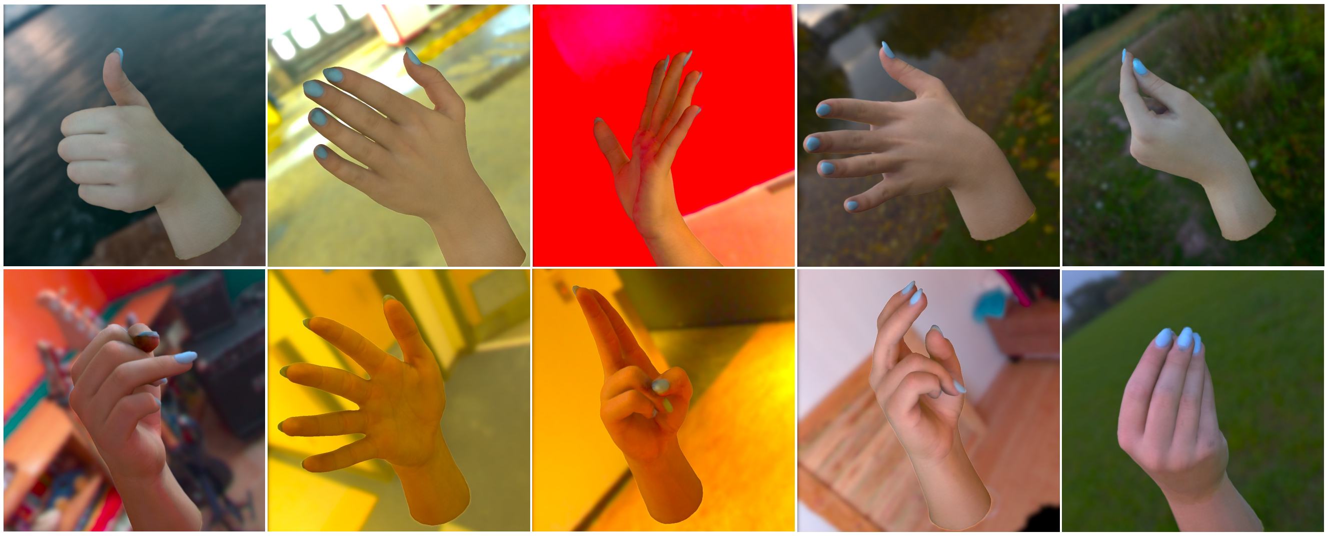

Figure 9. **Qualitative results of the student model with environment map lighting.** We sample the poses from the validation dataset of Subject 2.