Deep Reflectance Volumes: Relightable Reconstructions from Multi-View Photometric Images

Sai Bi[1], Zexiang Xu[1,2], Kalyan Sunkavalli[2], Miloš Hašan[2], Yannick Hold-Geoffroy[2], David Kriegman[1], Ravi Ramamoorthi[1]

[1] University of California, San Diego
[2] Adobe Research

1 BRDF Model

Essentially any differentiable BRDF model can be incorporated in our framework to model the appearance of real-world objects. In this paper we apply a version of the microfacet BRDF model proposed by Walter et al. [4], with simplifications introduced by Karis [1]. Let $\boldsymbol{\omega}_o$, $\boldsymbol{\omega}_i$ be the view and light direction, $\boldsymbol{n}$, $\boldsymbol{a}$, γ be the normal, diffuse albedo and roughness. Our BRDF model is defined as:

$$f(\boldsymbol{\omega}_o, \boldsymbol{\omega}_i, \boldsymbol{a}, \boldsymbol{n}, \gamma) = \frac{\boldsymbol{a}}{\pi} + \frac{D(\boldsymbol{h}, \boldsymbol{n}, r) F(\boldsymbol{\omega}_o, \boldsymbol{h}) G(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o, r)}{4(\boldsymbol{n} \cdot \boldsymbol{\omega}_i)(\boldsymbol{n} \cdot \boldsymbol{\omega}_o)} \tag{1}$$

where $D(\boldsymbol{h}, \boldsymbol{n}, r)$, $F(\boldsymbol{\omega}_o, \boldsymbol{h})$ and $G(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o, \boldsymbol{h}, r)$ are the *normal distribution*, *fresnel* and *geometric terms* respectively. These terms are defined as follows:

$$\begin{aligned}
D(\boldsymbol{h}, \boldsymbol{n}, \gamma) &= \frac{\alpha^2}{\pi \left[(\boldsymbol{n} \cdot \boldsymbol{h})^2 (\alpha^2 - 1) + 1 \right]^2} \\
\alpha &= \gamma^2 \\
F(\boldsymbol{\omega}_o, \boldsymbol{h}) &= F_0 + (1 - F_0) 2^{-[5.55473(\boldsymbol{\omega}_o \cdot \boldsymbol{h}) + 6.8316](\boldsymbol{\omega}_o \cdot \boldsymbol{h})} \\
G(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o, \gamma) &= G_1(\boldsymbol{\omega}_o, \boldsymbol{n}) G_1(\boldsymbol{\omega}_i, \boldsymbol{n}) \\
G_1(\boldsymbol{\omega}_o, \boldsymbol{n}) &= \frac{\boldsymbol{n} \cdot \boldsymbol{\omega}_o}{(\boldsymbol{n} \cdot \boldsymbol{\omega}_o)(1 - k) + k} \\
G_1(\boldsymbol{\omega}_i, \boldsymbol{n}) &= \frac{\boldsymbol{n} \cdot \boldsymbol{\omega}_i}{(\boldsymbol{n} \cdot \boldsymbol{\omega}_i)(1 - k) + k} \\
k &= \frac{(\gamma + 1)^2}{8}
\end{aligned}$$

where we set $F_0 = 0.05$ as suggested in [1]. Correspondingly, the final reflected radiance f_r in Eqn. 8 in the paper is computed as:

$$f_r(\boldsymbol{\omega}_o, \boldsymbol{\omega}_i, \boldsymbol{n}(\boldsymbol{x}_s), R(\boldsymbol{x}_s)) = f(\boldsymbol{\omega}_o, \boldsymbol{\omega}_i, \boldsymbol{a}(\boldsymbol{x}_s), \boldsymbol{n}(\boldsymbol{x}_s), \gamma(\boldsymbol{x}_s))(\boldsymbol{n}(\boldsymbol{x}_s) \cdot \boldsymbol{\omega}_i) \tag{2}$$

where $\boldsymbol{a}(\boldsymbol{x}_s)$ and $\gamma(\boldsymbol{x}_s)$ are the diffuse albedo and roughness at $\boldsymbol{x}_s$.

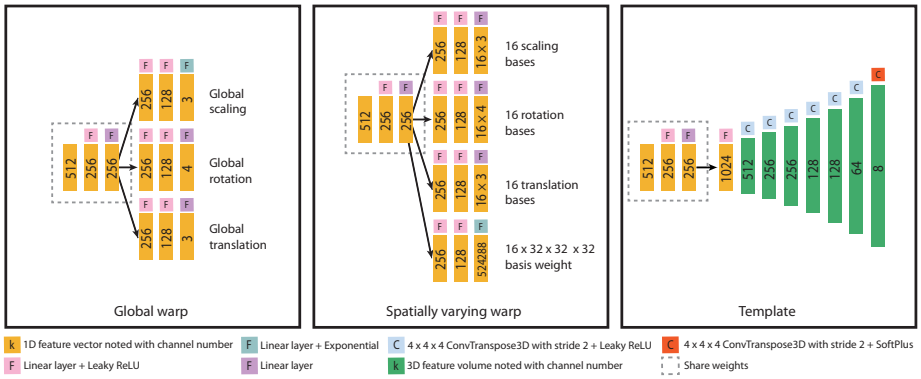

Fig. 1. Our network architecture.

	Pony	Girl	House	Disney	Animals	Captain
min	0.60	0.82	1.22	0.25	0.29	0.68
max	10.00	9.19	9.54	10.09	9.28	14.52
mean	5.35	7.66	5.83	7.25	6.49	6.92

Table 1. The minimum, maximum and average angles (in degrees) between the test views in the supplementary video and their nearest training views.

2 Network Architecture

Fig. 1 shows an overview of our network architecture. Our network starts from a 512-channel encoding vector initialized using random samples from a normal distribution. The encoding vector first goes through two fully connected layers and then is fed to different decoders to predict the global warping parameters, spatially varying warping parameters, and the template volume. The global warping parameters W_g consist of a 3-channel scaling vector, a 3-channel translation vector and a 4-channel rotation vector represented as a quaternion. The spatially varying parameters consist of 16 warping bases $\{W_j\}_{j=1}^{16}$ and a weight volume B. Similar to the global warping, each warping basis is composed of a scaling, a translation and a rotation. The weight volume B has 16 channels and a resolution of $32 \times 32 \times 32$, which encodes the spatially varying weight of each basis. Finally, the template volume V has a resolution of $128 \times 128 \times 128$; it has 8 channels with 1 channel for opacity, 3 channels for normal, 3 channels for diffuse albedo and 1 channel for roughness. We also transform the albedo and roughness to the range of $[0, 1]$ and normalize the predicted normal vectors.

3 Testing Specifications

In the supplementary video, we show renderings of the captured object under novel viewpoints and lighting. Note that our training images are captured with

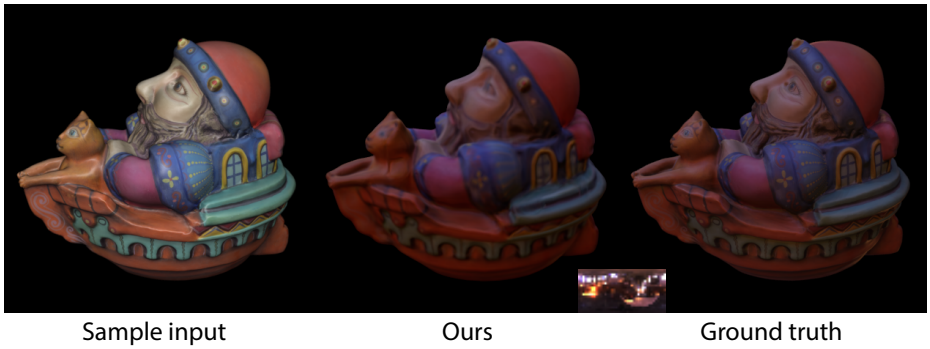

Fig. 2. Comparison with ground truth on relighting under environment illumination. The environment map used for rendering is shown at the bottom.

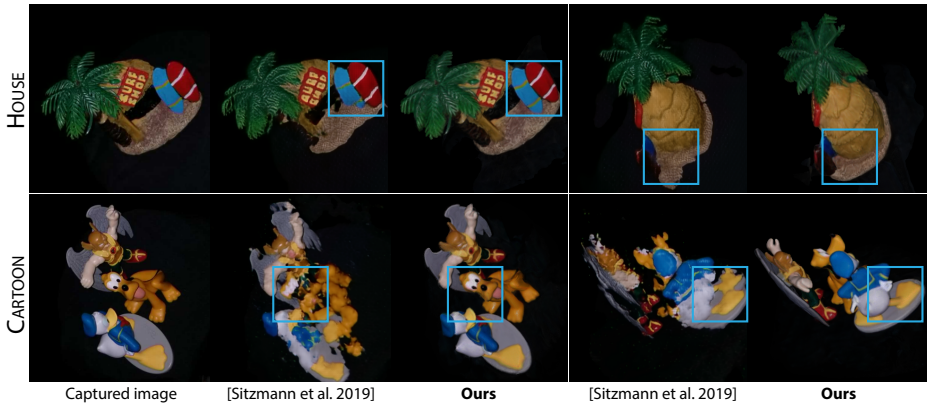

Fig. 3. Comparison against Sitzmann et al. [3] on synthesizing novel views under collocated lights. Our method is able to generate high-quality results with fewer artifacts.

collocated light and camera, and the relighting results in the video demonstrate that our volumetric representation can generalize to novel lighting conditions. In Tab. 1, we report the minimum, maximum and average angles between the test views in the video and their nearest training views. Such a large angle difference also shows that our deep reflectance volume generalizes well to novel views.

4 Results on Synthetic Data

In addition to the real captures, we also evaluate our method on a synthetic dataset where we render a synthetic scene from multiple viewpoints under collocated camera and light. We compare our view synthesis and relighting results with the ground truth renderings. Please check the supplementary video for comparisons.

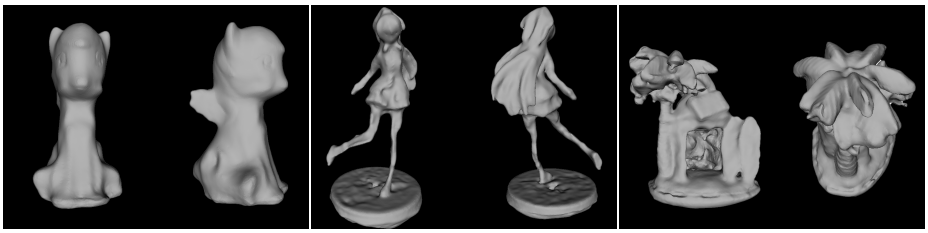

Fig. 4. Geometry reconstructed from Nam et al. [2].

By linearly combining the relit images under each light corresponding to pixels of an environment map, our method also supports rendering of the scene under novel environment illumination. In Fig. 2 we demonstrate our environment map relighting result and compare it to the ground truth renderings with a physically-based renderer. From the figure we can see that our method can generate visually plausible results.

5 Comparison on View Synthesis

In Fig. 3 we show a visual comparison against the method proposed by Sitzmann et al. [3] on synthesizing novel views under collocated lights. Sitzmann et al. learn a 3D-aware neural representation to encode the view-dependent appearance of captured scenes. Their method cannot model the complex geometry and appearance of our real scenes. As we can see from the result, Sitzmann et al. cannot synthesize novel views correctly and generates distorted images with undesired structures. In contrast, our method is able to produce images of much higher quality.

6 Mesh-Based Appearance Acquisition

In Fig. 4 we show the optimized geometry from Nam et al. [2]. They leverage the state-of-the-art multi-view stereo (MVS) framework to get an initial geometry, and further perform an optimization to refine it; however they still fail to recover the faithful geometry for such challenging scenes where there are textureless and thin-structured regions, thus resulting in degraded quality in reproduced appearance, as shown in the supplementary video.

References

1. Karis, B.: Real shading in unreal engine 4. Proc. Physically Based Shading Theory Practice (2013)
2. Nam, G., Lee, J.H., Gutierrez, D., Kim, M.H.: Practical SVBRDF acquisition of 3D objects with unstructured flash photography. In: SIGGRAPH Asia 2018. p. 267. ACM (2018)
3. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2437–2446 (2019)
4. Walter, B., Marschner, S.R., Li, H., Torrance, K.E.: Microfacet models for refraction through rough surfaces. Rendering techniques (2007)