Supplementary Material for Self-calibrating Photometric Stereo by Neural Inverse Rendering

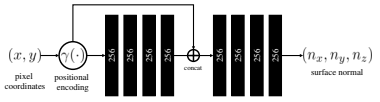
Junxuan Li and Hongdong Li

Australian National University

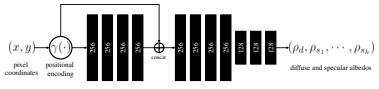
1 Implementation Details

Network architectures Here, we describe our network architectures in detail. In Fig. S1a, the surface normal net $N_\Theta(\cdot)$ uses 8 fully-connected layers with 256 channels. It takes positional encoded [11] pixel coordinates $\gamma(x), \gamma(y)$ as input, directly output the surface normal at that position $\boldsymbol{n} = [n_x, n_y, n_z]^T$. As shown in Fig. S1b, the material net $M_\Phi(\cdot)$ uses the same structure but with 3 more fully-connected layers than normal net. It takes the same positional encoded pixel coordinates input and outputs the diffuse and specular albedos of the surface point. As shown in Fig. S1c, the lighting network $L_\Psi(\cdot)$ consists of 7 convolutional ReLU layers and 3 fully connected layers. It takes the image with size $H \times W \times 3$ as input, directly outputs the light intensity e and light direction $\boldsymbol{l}$ of that image.

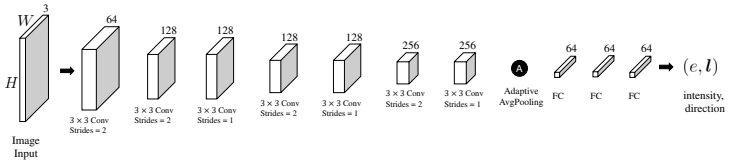
Early supervision Following previous works [2, 8], we additionally use the surface smoothness constraints and shape-from-contour priors as the early supervision in our network. After early-stage training in the first half iterations, we discard these priors and train the network via photometric loss.



(a) Normal network.



(b) Material network.



(c) Lighting network.

Fig. S1: The network architectures of our networks.

2 Visualization of the effectiveness of PSB

Recall that, although the GBR ambiguity can be reduced up to a binary concave/convex ambiguity under our model, there is no guarantee that no local minima exist during the optimization. To effectively avoid the local minimas during the optimization, we propose the progressive specular bases (PSB) for the network. In Fig. S2, we provide the visual comparison between the model with PSB and the model without PSB. The first row displays the observed ground truth image under a light source, the ground truth light distribution, and the ground truth surface normal. The second row and third row display the reconstructed image, the estimated light distribution, the estimated normal, the error map of estimated normal, and the estimated shape from our "with PSB model" and "without PSB model" respectively.

As we can see, the "without PSB" produces a worse light and normal estimation. Both the light and normal are "shifted" along the z axis. However, its reconstructed image still presents a similar quality to the observed ground truth (PSNR: 40.06dB). This observation coincides with the observation from Belhumeur [3], where they also observed that the differences in shape are hard to be discerned from the frontal images given a small scale along the z axis.

The PSB can provide prior information to the network and limit the space of possible solutions by forcing the network to fit on the shiny specularities first in the early stage of optimization. Hence, by applying with the PSB, even with a poor network initialization, our network can still effectively avoid the local minimas to achieve better results.

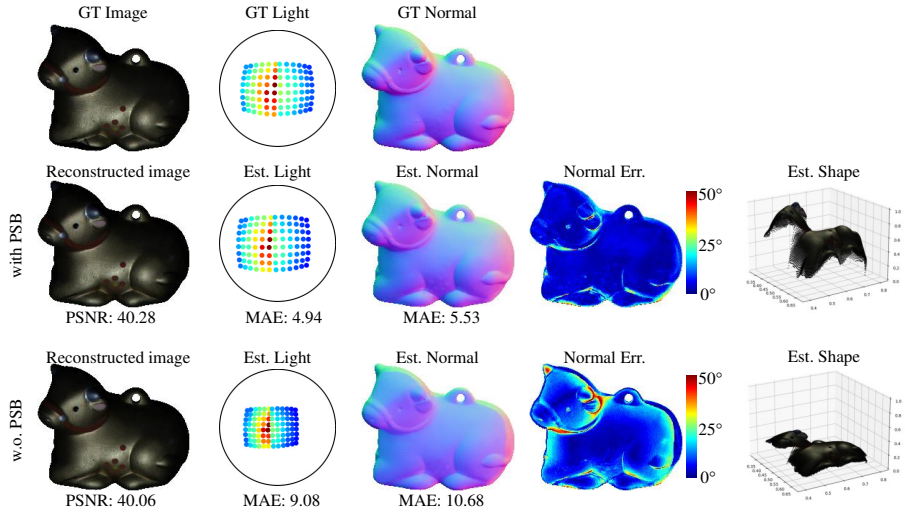


Fig. S2: Visualized comparisons of with/without using the *progressive specular basis* (PSB).

3 Ablation study on lighting model

In this section, we conduct two experiments to showcase the effectiveness of the lighting network. As shown in Fig. S3: on the first row, we showcase the observed(GT) image, ground truth lights and normals; on the second row, we display reconstructed image and estimated result using the model with lighting network $L_\Psi(\cdot)$; on the third row, we present results using the model without lighting network and takes randomized lights as initialization.

Without using the lighting network, we take randomized lights as initialization. Our network may sometimes produce a flipped surface as a result, as shown in the third row in Fig. S3. As we can see in the second row and third row in Fig. S3, the estimated lights and normals are flipped in the x, y axis. In the third row, the mean angular error (MAE) for light direction is 55.47 degrees, and normal error is 91.07 degrees. However, its reconstructed image is almost identical to the observed ground truth image.

During the experiments, we observed that this convex/concave ambiguity can be easily resolved by providing the model with a coarse lighting estimation, as shown in second row in Fig. S3. Our lighting model $L_\Psi(\cdot)$ can provide a coarse lighting estimation as the starting point, which is sufficient to for the followed self-supervised network to further refine the coarse results and produce the correct lights.

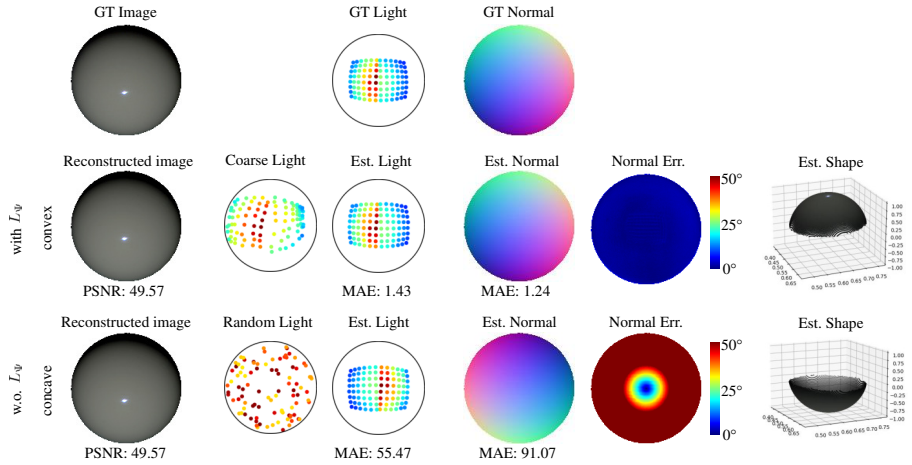


Fig. S3: Visualization of the effectiveness of lighting model.

4 Robustness on Sparse Inputs

In this section, we present the results on DiLiGenT [12] dataset with only 16 images at the inputs. Following previous works on sparse inputs for photometric stereo [9], we selected 16 images as input for our method and others for comparison. The errors are shows in Tab. S1. As we can see from the table, our method still outperform the state-of-the-art with only 16 images. Besides, with 16 images as input, our method only drop 0.72 degrees in MAE in normal estimation, while GCNet[7]+PSFCN[6] drops 2.04 degrees in MAE. It also demonstrate that our method is robust against sparse input.

Table S1: Quantitative comparison on DiLiGenT with only 16 images at input.

(a) MAE of surface normal.

model	All images	16 images
Ours	7.05	7.77
GCNet[7]+PSFCN[6]	8.70	10.74

(b) Scale-invariant relative error of light intensities.

model	All images	16 images
Ours	0.0365	0.0548
GCNet[7]	0.0519	0.0550

(c) MAE of light directions.

model	All images	16 images
Ours	4.02	5.02
GCNet[7]	3.32	4.04

5 Results on DiLiGenT benchmark

In this section, we present the results on DiLiGenT [12] dataset, as shown in the following Fig. S6, S7 and S8. For each object, the first row displays the ground truth lighting, ours estimated lighting, and lighting results from GCNet [7] and SDPS-Net [5]. The second row displays the ground truth surface normal and estimated surface normal by ours and competing methods. The last row display the observed image and the error map of the estimated surface normal. We also present the quantitative evaluation for lighting and normal below the lighting and error map. Note that UPS-FCN [6] can not estimate the lighting.

Results on almost Lambertian surface As we can see from the results, our method works well for specular objects, as well as objects that appear to be very diffuse, such as "Cat". In order to better understand why our method also works well on objects like "Cat", we visualized the reconstructed terms ρ_d and ρ_s in Fig. S4. Figure S4 shows that the "Cat" is not purely diffuse and contains very soft specularities. Our method is able to capture and use these soft specularities as clues for estimating the surface normal.

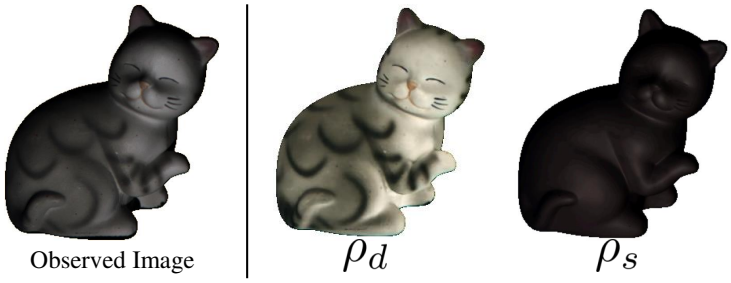


Fig. S4: Visualization of reconstructed ρ_d and ρ_s in object "Cat".

6 Results on Apple&Gourd dataset

In this section, we present the results on Apple&Gourd [1] dataset. In Fig. S9, for each object, the first row displays the ground truth lighting, ours estimated lighting, and lighting results from GCNet [7]. The second row displays the observed image and estimated surface normal by ours and competing methods. Note that there is no ground truth surface normal available in this dataset, so we only visualized compare the normal results. As shown in "Gourd2", it is clear that our estimated normal present higher quality than previous state-of-the-art method GCNet [7]+PSFCN [6].

7 Results on synthetic dataset with 100 MERL BRDFs

To evaluation our method across different surface materials and BRDFs, we test our method on a publicly available synthetic dataset[1]: GCNet-Synthetic [7]. The dataset consists of two rendered synthetic objects: Dragon and Armadillo for testing. This dataset was rendered with 100 MERL [10] BRDFs under 82 random light directions using physically based renderer Mitsuba[2].

We showcase the results in Fig. S10 and Fig. S11. As we can see from the figures, our method produce comparable results to GCNet [7].

[1] https://github.com/guanyingc/UPS-GCNet
[2] http://mitsuba-renderer.org/

We dive into the MERL dataset and found that our method fails to fit the materials such as "steel", "chrome", and "chrome-steel", where they generally present asymmetric highlights as shown in Fig. S5. Prior work [4] believed that these anomaly asymmetric highlights could be caused by the lens flare. We believe that using a different BRDF model to account for these effects can improve the performance on these materials. We are happy to consider this as a future direction.

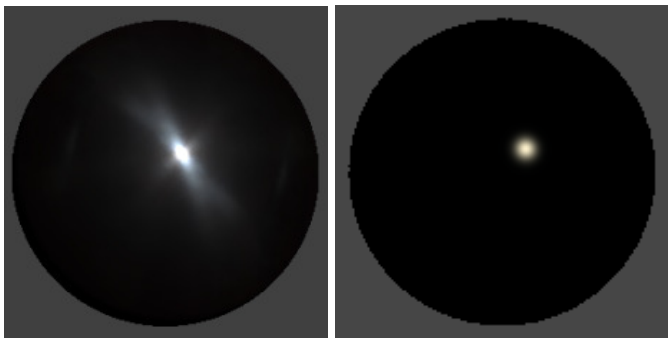


Fig. S5: Rendered sphere of "steel". Left is the data from MERL. Right is our estimated result.

8 Self-captured images outside of the laboratory

We captured 55 images with a Nikon camera and a handheld flashlight. The target object is captured in a regular livingroom environment with lights off. The captured image and our estimated results (normals, shadings, and lights) are shown in Fig. S12. As we can see from the results, our method still performs very well in a non-laboratory environment.

9 Future works

We believe that our method, with some adaptations, can be extended to solve the problem under many other assumptions, such as specularity detection, multi-view photometric stereo, photometric stereo under multi-light-sources and natural illumination. Our method inverse renders the object to shapes and materials. Hence, the specularity detection is also available at output, as shown in Fig. S4. A possible adaptation for multi-view photometric stereo is to apply our algorithm to each view of the object, and then fuse the normal map from different views to obtain the full geometry. We can also model the environment map as Spherical-Gaussians to enable fast integration of BRDF and lighting in natural-illumination and multi-light-sources.

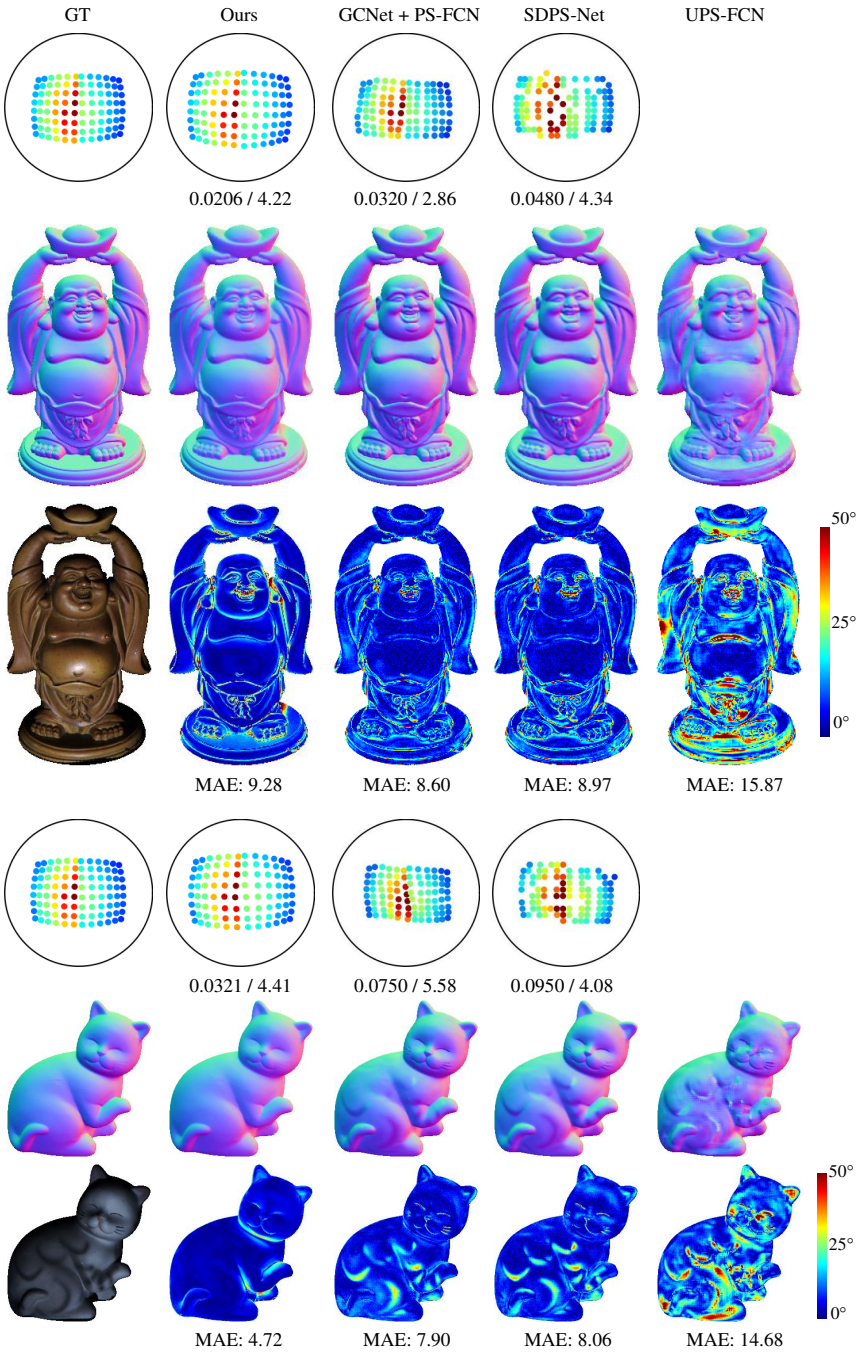


Fig. S6: Results for "Buddha" and "Cat" from DiLiGenT dataset.

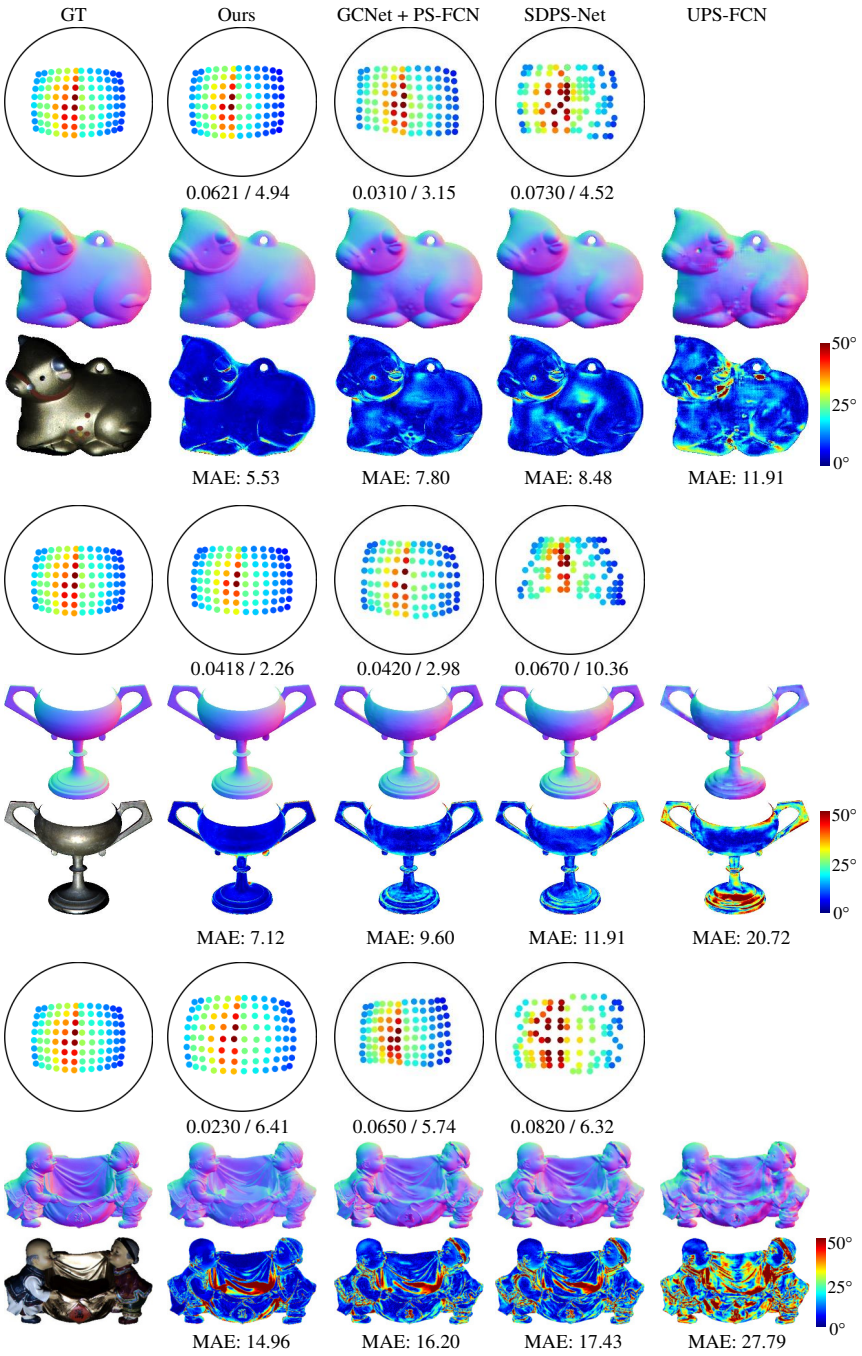


Fig. S7: Results for "Cow" , "Goblet", and "Harvest" from DiLiGenT dataset.

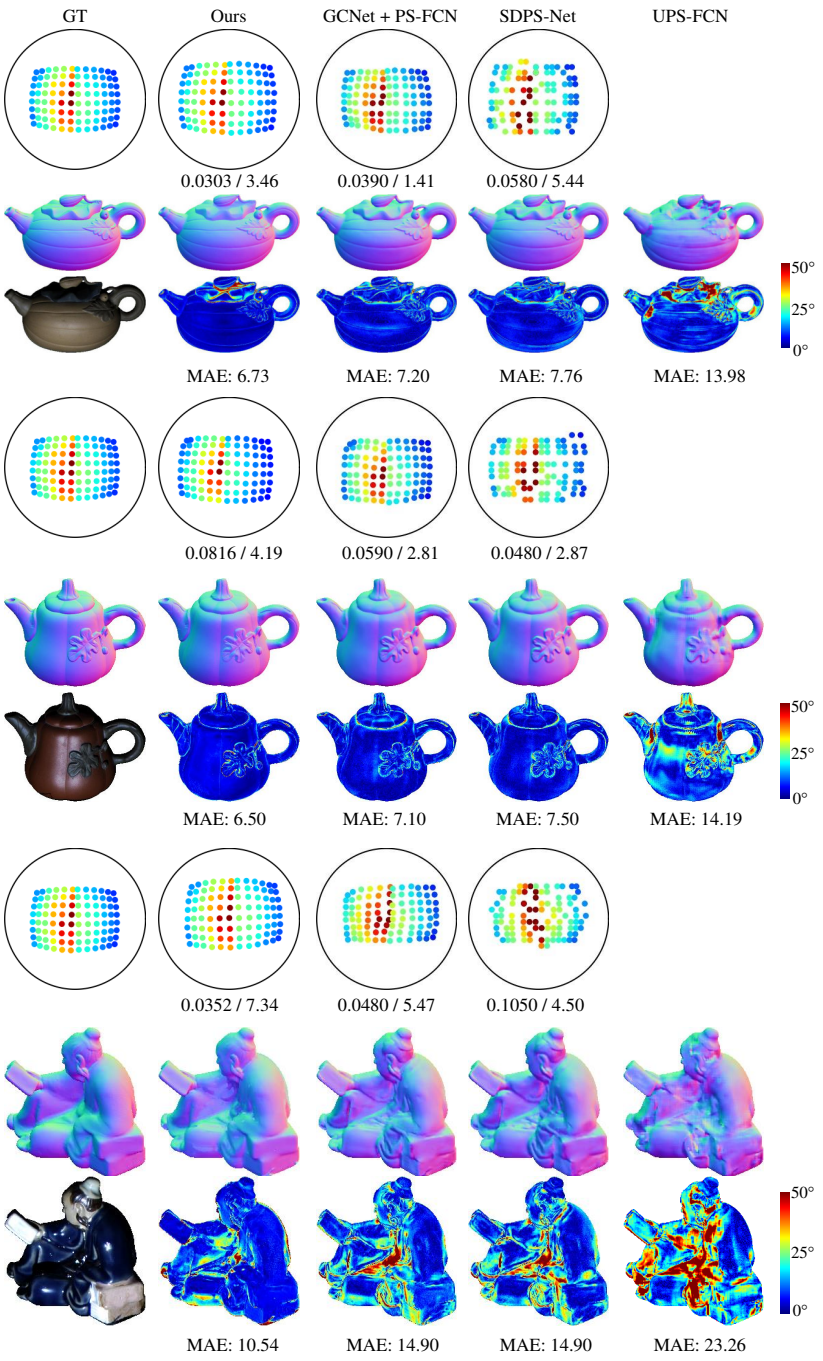


Fig. S8: Results for "Pot1" , "Pot2", and "Reading" from DiLiGenT dataset.

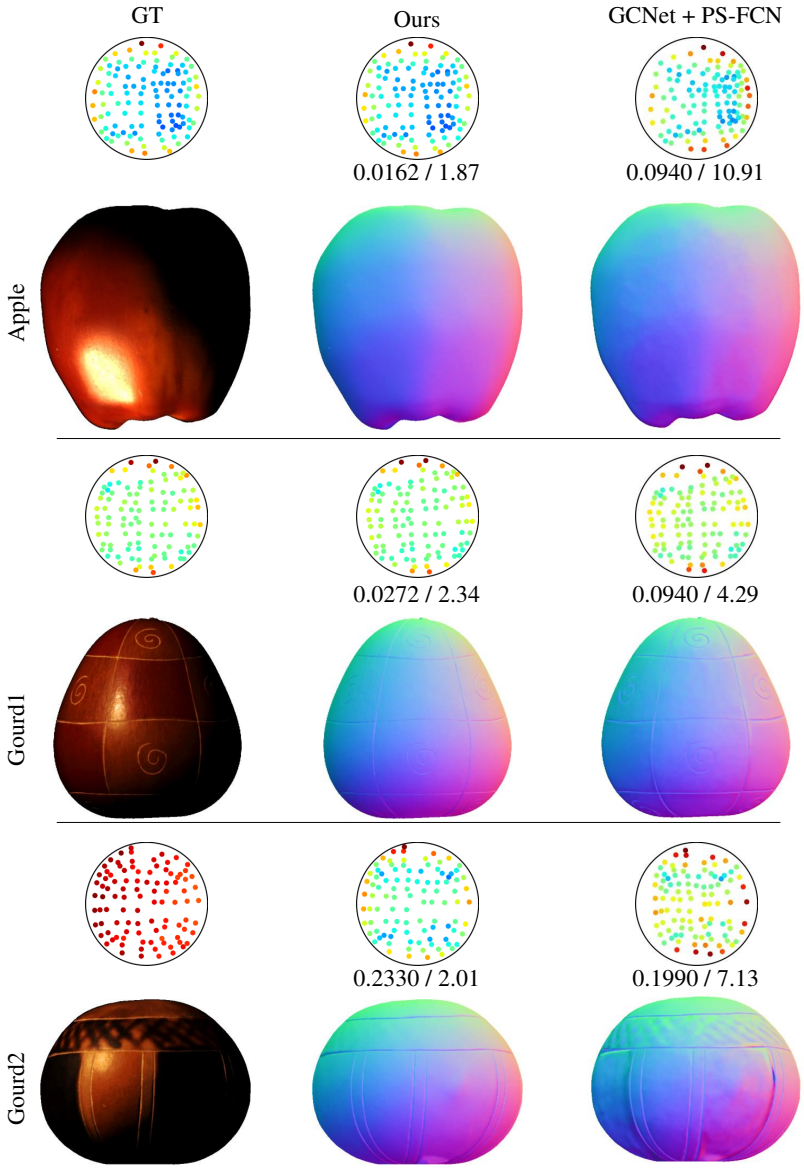
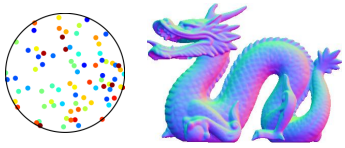
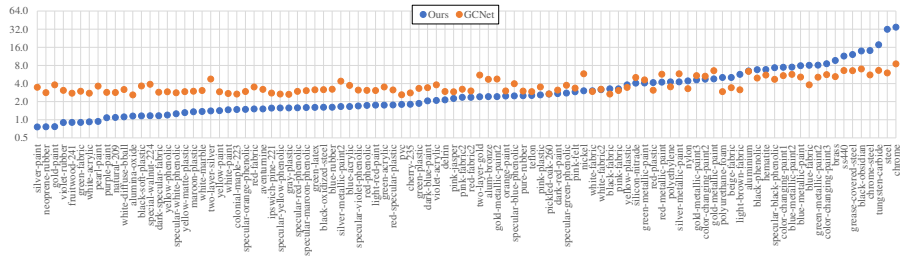


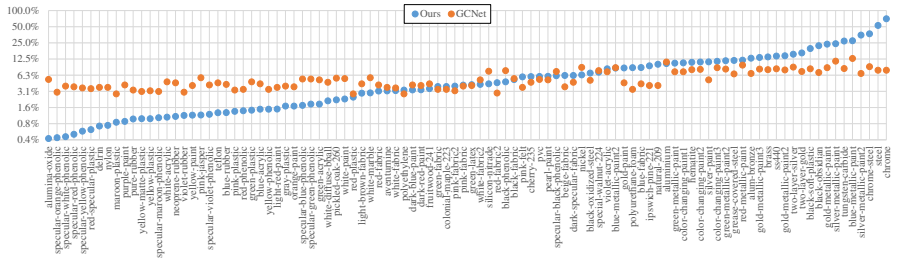
Fig. S9: Results for "Apple" , "Gourd1", and "Gourd2" from Apple&Gourd dataset.



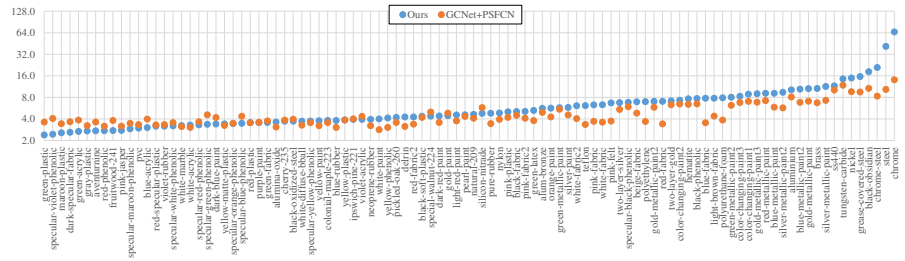
(a) Ground truth of lights and surface normals.



(b) MAE of light directions.

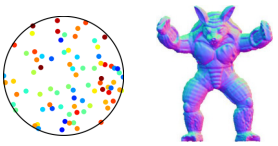


(c) Scale-invariant relative error in percentage of light intensities.

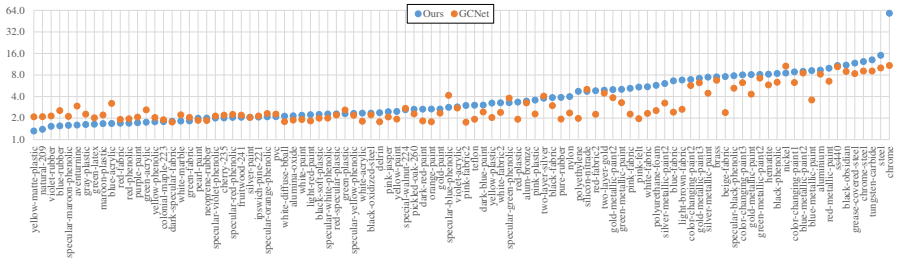


(d) MAE of surface normals.

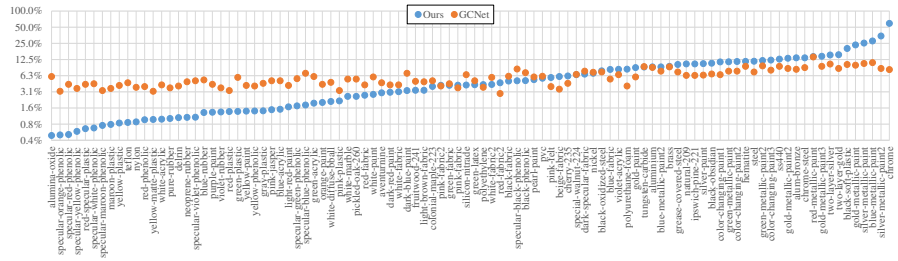
Fig. S10: Comparison on object "Dragon" rendered with 100 MERL BRDFs.



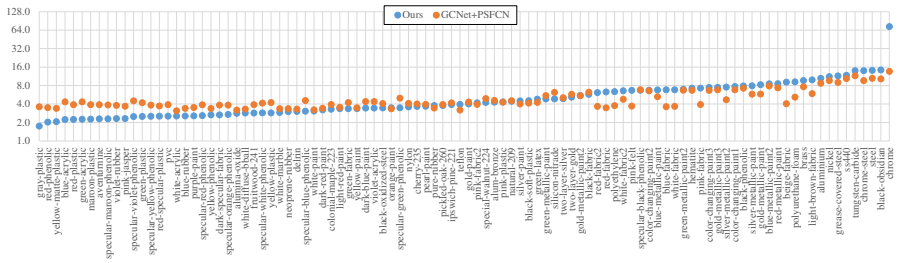
(a) Ground truth of lights and surface normals.



(b) MAE of light directions.



(c) Scale-invariant relative error in percentage of light intensities.



(d) MAE of surface normals.

Fig. S11: Comparison on object "Armadillo" rendered with 100 MERL BRDFs.

Fig. S12: The captured image of "CokeCan" and our estimations.

References

1. Alldrin, N., Zickler, T., Kriegman, D.: Photometric stereo with non-parametric and spatially-varying reflectance. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
2. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE transactions on pattern analysis and machine intelligence **37**(8), 1670–1687 (2014)
3. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. International journal of computer vision **35**(1), 33–44 (1999)
4. Burley, B., Studios, W.D.A.: Physically-based shading at disney. In: ACM SIGGRAPH. vol. 2012, pp. 1–7. vol. 2012 (2012)
5. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Self-calibrating deep photometric stereo networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8739–8747 (2019)
6. Chen, G., Han, K., Wong, K.Y.K.: Ps-fcn: A flexible learning framework for photometric stereo. In: European Conference on Computer Vision. pp. 3–19. Springer (2018)
7. Chen, G., Waechter, M., Shi, B., Wong, K.Y.K., Matsushita, Y.: What is learned in deep uncalibrated photometric stereo? In: European Conference on Computer Vision. pp. 745–762. Springer (2020)
8. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16221–16230 (2022)
9. Li, J., Robles-Kelly, A., You, S., Matsushita, Y.: Learning to minify photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7568–7576 (2019)
10. Matusik, W., Pfister, H., Brand, M., McMillan, L.: A data-driven reflectance model. ACM Transactions on Graphics **22**(3), 759–769 (Jul 2003)
11. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)
12. Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)