Supplementary Material for "What is Learned in Deep Uncalibrated Photometric Stereo?"

Guanying Chen[1] Michael Waechter[2] Boxin Shi[3,4] Kwan-Yee K. Wong[1] Yasuyuki Matsushita[2]
[1]The University of Hong Kong [2]Osaka University [3]Peking University [4]Peng Cheng Laboratory

In this supplemental material,

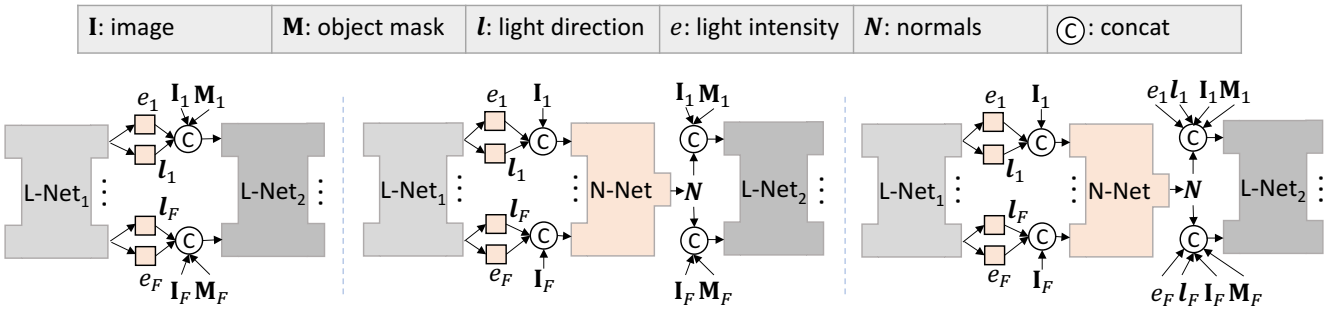
1. we show more feature visualizations of LCNet [2] in the supplementary video,
2. we provide more discussions regarding the proposed cascade structure,
3. we present detailed results of GCNet on the synthetic dataset rendered with 100 MERL BRDFs [6], and
4. we compare our GCNet with previous state-of-the-art uncalibrated methods on 5 real datasets, including DiLiGenT benchmark [8], DiLiGenT test dataset [8], Light Stage Data Gallery [4], Harvard photometric stereo dataset [10], and Gourd&Apple dataset [1].

Contents

1. Feature visualization of LCNet

In our supplementary video[1], we show more feature visualizations of LCNet [2] beyond the three features shown in the paper.

2. More discussions regarding the network architecture



(a) L-Net$_1$ + L-Net$_2$ (b) L-Net$_1$ + N-Net + L-Net$_2^{\text{(w/o shading)}}$ (c) L-Net$_1$ + N-Net + L-Net$_2^{\text{(w/o shading; w/ light)}}$

Figure S1. Three different cascaded structures. We omit the input of L-Net$_1$ and the output of L-Net$_2$ for all models to simplify the illustration

Table S1. Lighting estimation results on the synthetic test dataset. The results are averaged over 100 MERL BRDFs (bold fonts indicates the best).

ID	model	*Sphere*		*Bunny*		*Dragon*		*Armadillo*	
		direction	intensity	direction	intensity	direction	intensity	direction	intensity
0	L-Net$_1$ + L-Net$_2$	2.92	0.051	4.37	0.058	5.99	0.079	5.31	0.077
1	L-Net$_1$ + N-Net + L-Net$_2^{\text{(w/o shading)}}$	2.79	**0.046**	3.21	0.056	4.63	0.072	4.29	0.062
2	L-Net$_1$ + N-Net + L-Net$_2^{\text{(w/o shading; w/ light)}}$	2.83	0.047	3.50	**0.051**	5.36	0.075	4.04	0.068
3	L-Net$_1$ + N-Net + L-Net$_2$	**2.52**	0.052	**2.90**	0.054	**4.20**	**0.061**	**3.92**	**0.060**

Comparison of different cascaded structures Cascaded structure is a popular strategy to improve performance. Ours, however, was no naïve extension of LCNet, but carefully designed based on our analysis. We compared our structure with three different structures (see Fig. S1) to verify our method's effectiveness.

Figure S1 (a) is a common structure to refine a network's estimation using another similar network. As discussed in the paper, L-Net's bottleneck is the lack of inter-image information (*e.g.*, normals) during feature extraction, making this structure sub-optimal (compare the experiments with IDs 0 & 3 in Table S1). Figure S1 (b) is a sequential structure where L-Net$_2$ additionally takes estimated normals as input. However, the experiments with the IDs 1 & 3 show that taking the estimated shading (intra-image information) as input is beneficial. In Fig. S1 (c), L-Net$_2$ takes estimated normals and lightings as input. Our experiment shows that taking lightings as input can lead to faster convergence, but the final performance is worse than the proposed method, as show in the experiments with the IDs 2 & 3. We suspect that L-Net$_2$ becomes more dependent on input lightings if directly taking them as input during training. When L-Net$_1$'s estimated lightings are not accurate, the refined estimation may not be good.

We further compared the proposed structure with these three baseline structures on *Dragon* rendered with SVBRDFs. Specifically, we rendered 100 test objects by randomly sampling two MERL BRDFs and blending the BRDFs for the challenging *Dragon* using the material maps "ramp" and "irregular" shown in Table 6 of the paper. Table S2 shows that the proposed structure consistently outperforms three baseline structures on *Dragon* rendered with SVBRDFs, demonstrating the effectiveness of the proposed structure.

[1]https://guanyingc.github.io/UPS-GCNet

Table S2. Lighting estimation results on *Dragon* rendered with SVBRDFs.

model	uniform		ramp		irregular	
	direction	intensity	direction	intensity	direction	intensity
(a) L-Net$_1$ + L-Net$_2$	5.99	0.079	5.64	0.066	6.99	0.090
(b) L-Net$_1$ + N-Net + L-Net$_2^{\text{(w/o shading)}}$	4.63	0.072	4.72	0.060	6.44	0.082
(c) L-Net$_1$ + N-Net + L-Net$_2^{\text{(w/o shading; w/ light)}}$	5.36	0.075	5.39	0.059	5.34	0.077
proposed: L-Net$_1$ + N-Net + L-Net$_2$	**4.20**	**0.061**	**4.17**	**0.051**	**5.02**	**0.070**

2.1. Multiple cascaded structures

We also trained a network which appends an additional cascade structure of N-Net$_2$ and L-Net$_3$ after GCNet. The additional structure is trained step by step. We observed that this network performs slightly better on synthetic data, but worse on real data. This might be due to the increasing difficulty of training or an over-fitting problem.

2.2. Cyclic prediction strategy

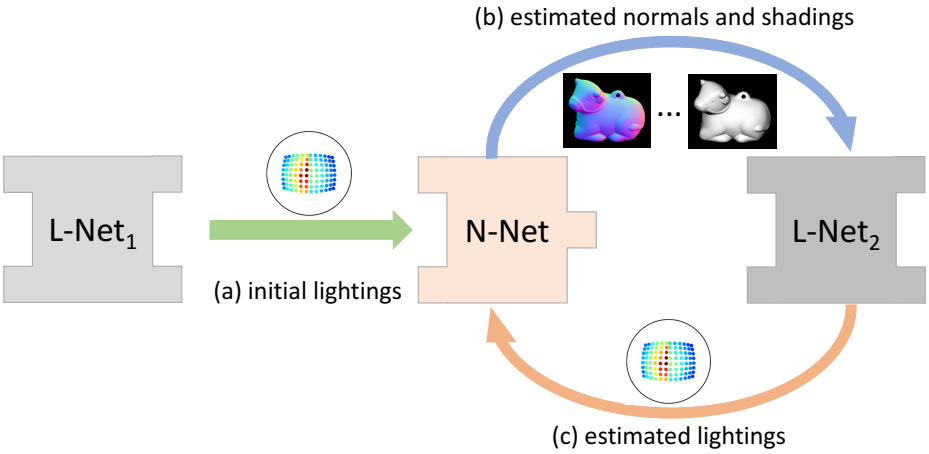


Figure S2. Illustration of the cyclic prediction strategy for GCNet.

The proposed GCNet first estimates initial lightings (L-Net$_1$) for surface normal estimation (N-Net), and then estimates better lightings based on the estimated surface normals and shadings (L-Net$_2$). A seemingly feasible idea to further improve results might be to iteratively estimate surface normals and lightings using N-Net and L-Net$_2$.

To verify this cyclic prediction strategy, at test time, we iteratively used the estimated lightings of L-Net$_2$ as input for N-Net, and the normals and shading of N-Net as input for L-Net$_2$ (see the cycle in Fig. S2). L-Net$_1$, N-Net, and L-Net$_2$ are trained as described in the paper, and the weights are fixed during testing.

Figure S3 shows the lighting estimation results of GCNet using the cyclic prediction strategy on five objects from the DiLiGenT benchmark. We can see that, although the errors converge within 20 iterations, the results do not always improve. For example, compared to the results obtained without cyclic prediction, light direction estimation on *Harvest* improved while it worsened on the *Ball*. The reason might be that the input data distributions of the networks (*i.e.*, N-Net and L-Net$_2$) are slightly shifted after each iteration and might therefore become different from those encountered during training. There is no guarantee that the result of GCNet improves through the cyclic prediction strategy.

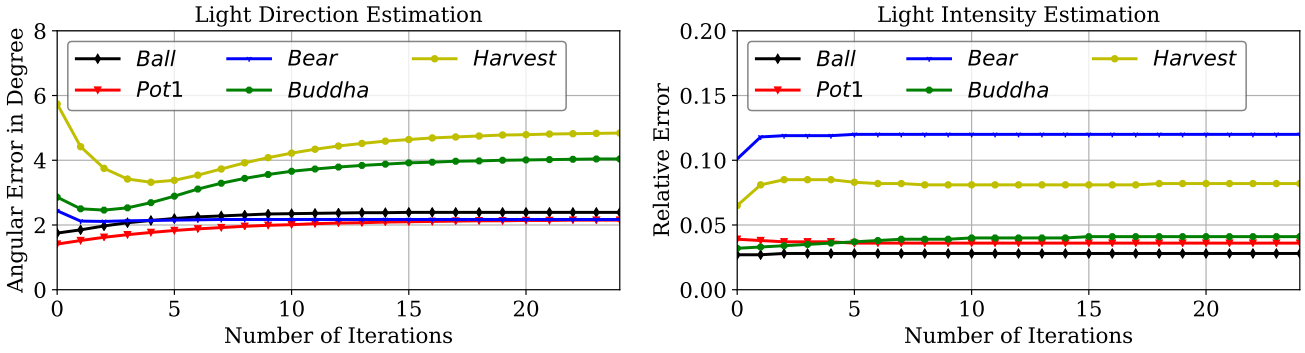
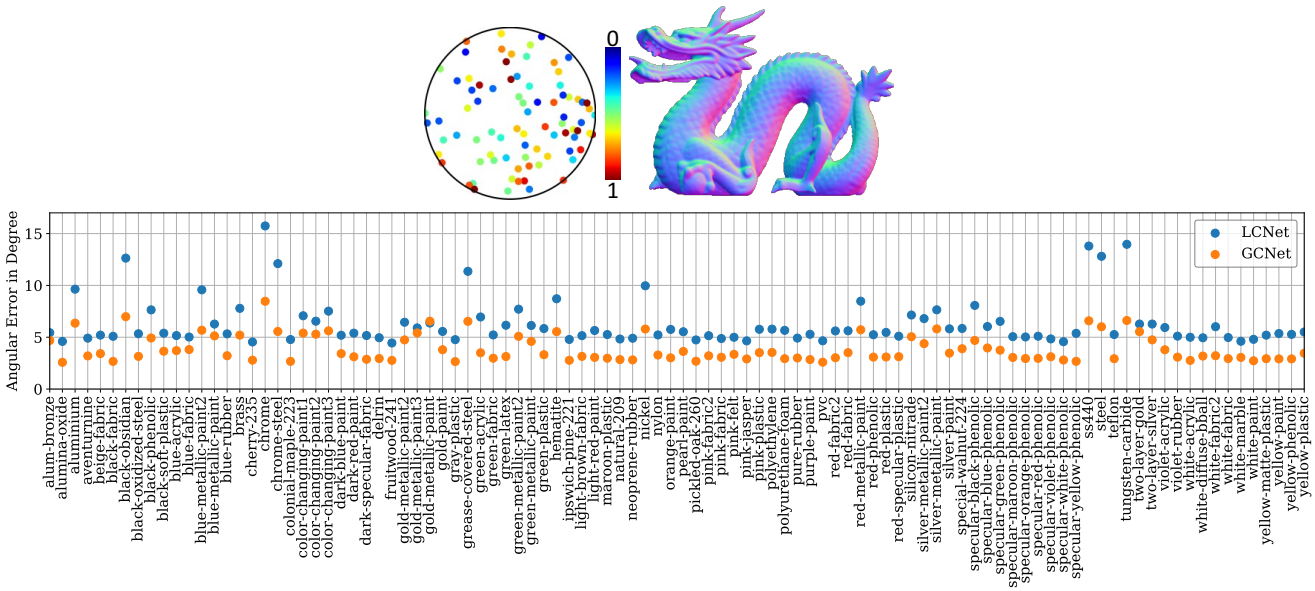


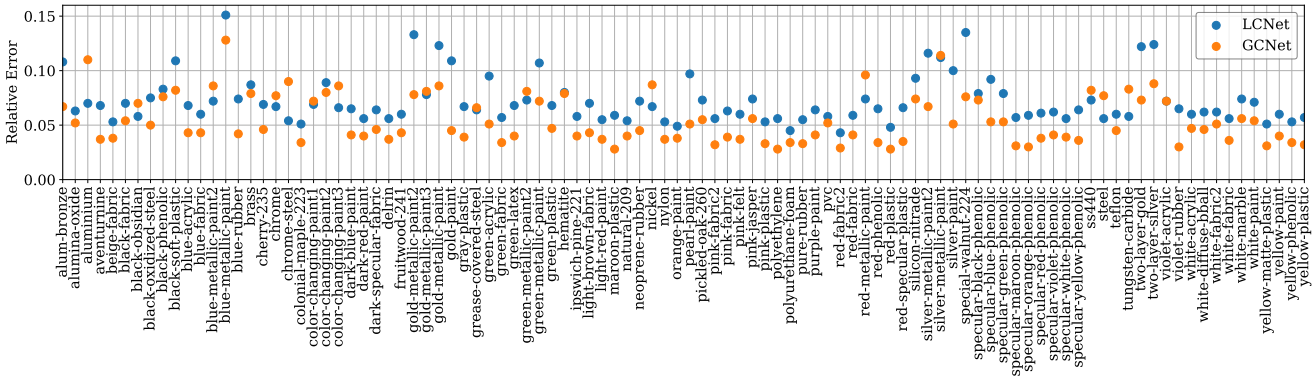
Figure S3. Lighting estimation results of GCNet on the DiLiGenT benchmark using the cyclic prediction strategy. Results at iteration 0 (*i.e.*, without cyclic prediction) are the results of GCNet proposed in our paper.

3. Detailed results on synthetic dataset rendered with 100 MERL BRDFs

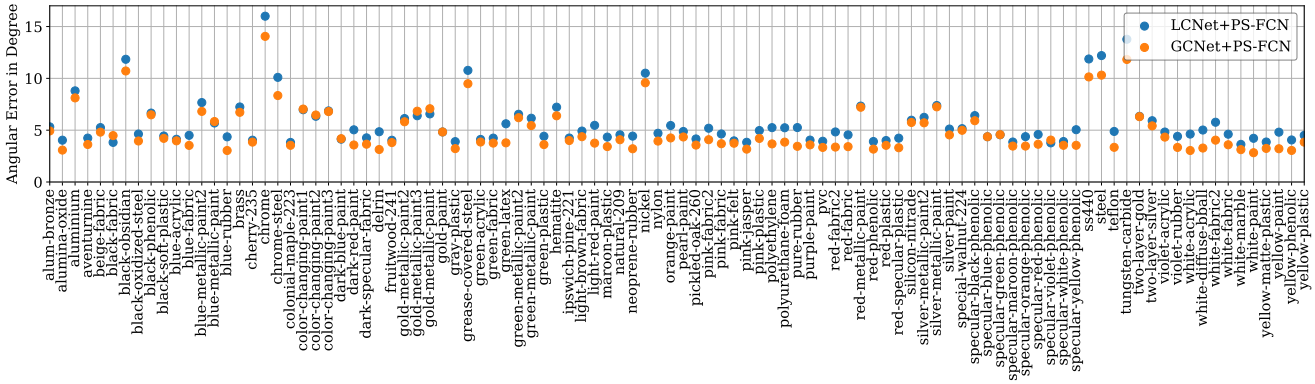
Figures S4 and S5 show detailed results of our method and LCNet [2] on the *Dragon* and *Armadillo* scenes rendered with 100 MERL BRDFs [6]. Our method clearly outperforms LCNet.



(a) Light direction estimation results. The average MAE over 100 BRDFs is 6.30 for LCNet and 3.88 for GCNet.

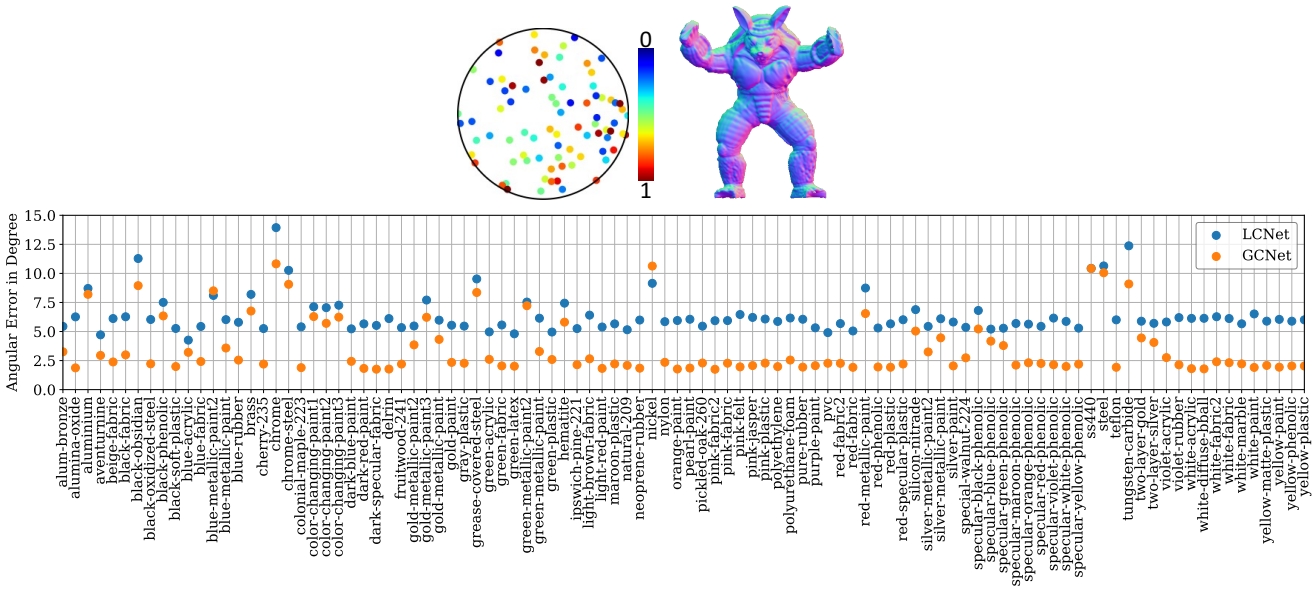


(b) Light intensity estimation results. The average relative error over 100 BRDFs is 0.072 for LCNet and 0.055 for GCNet.

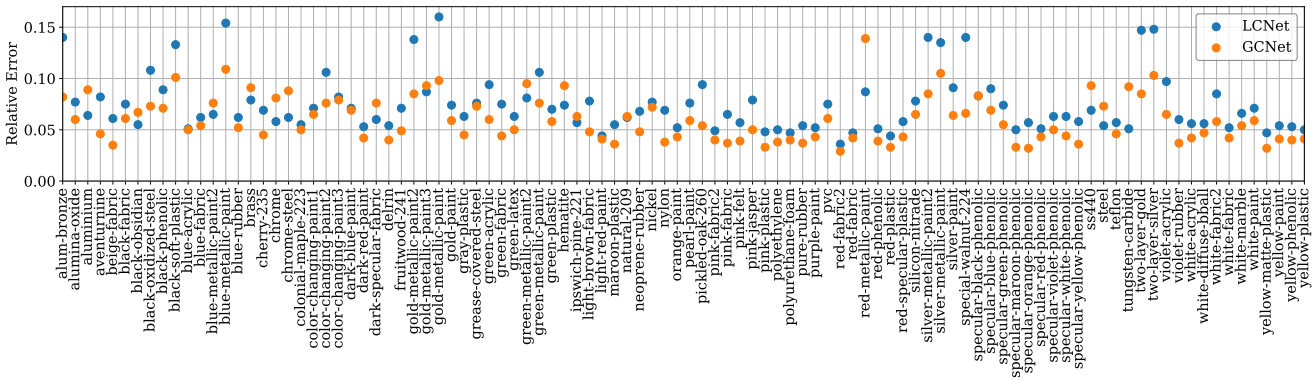


(c) Surface normal estimation results. The average MAE over 100 BRDFs is 5.59 for LCNet and 4.85 for GCNet.

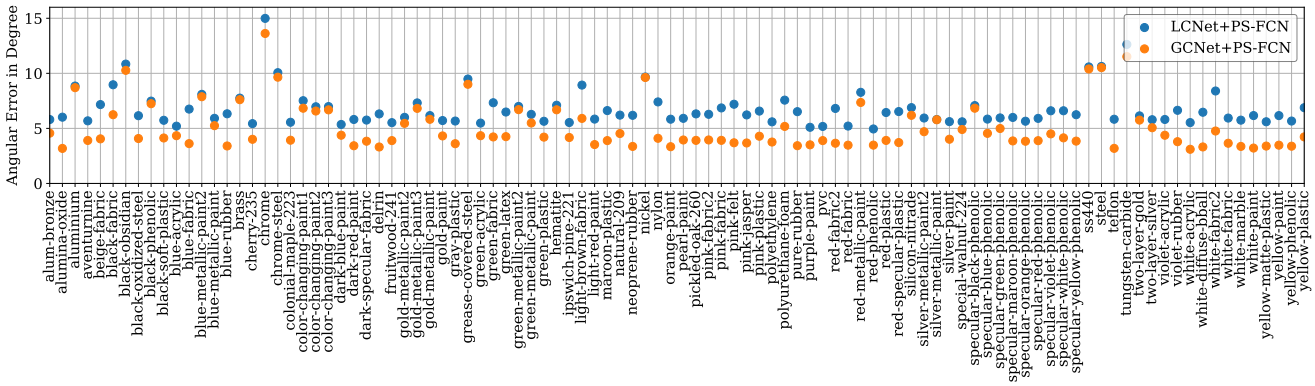
Figure S4. Comparison between LCNet [2] and GCNet on *Dragon* rendered with 100 MERL BRDFs.



(a) Light direction estimation results. The average MAE over 100 BRDFs is 6.37 for LCNet and 3.52 for GCNet.



(b) Light intensity estimation results. The average relative error over 100 BRDFs is 0.074 for LCNet and 0.060 for GCNet.



(c) Surface normal estimation results. The average MAE over 100 BRDFs is 6.73 for LCNet and 5.01 for GCNet.

Figure S5. Comparison between LCNet [2] and GCNet on *Armadillo* rendered with 100 MERL BRDFs.

4. Results on the DiLiGenT benchmark

In this section we show results on objects from the DiLiGenT Benchmark [8]. The DiLiGenT benchmark contains 10 real objects with different shapes and materials. Each object is captured under 96 different light directions and the calibrated lightings and ground-truth surface normals are provided for quantitative evaluation.

For each test object, *row 1* shows the ground-truth lightings and estimated lightings, *row 2* shows the ground-truth and estimated surface normals, and *row 3* shows an image of the object and error maps of the estimated surface normals. The values below the estimated lightings indicate the MAE of light directions and the relative error of light intensities. The values below the error maps indicate the MAE for the surface normal estimates. Note that UPS-FCN [3] only estimates the surface normals, and the normal estimation result of WT13 [9] was obtained from the DiLiGenT benchmark [8].

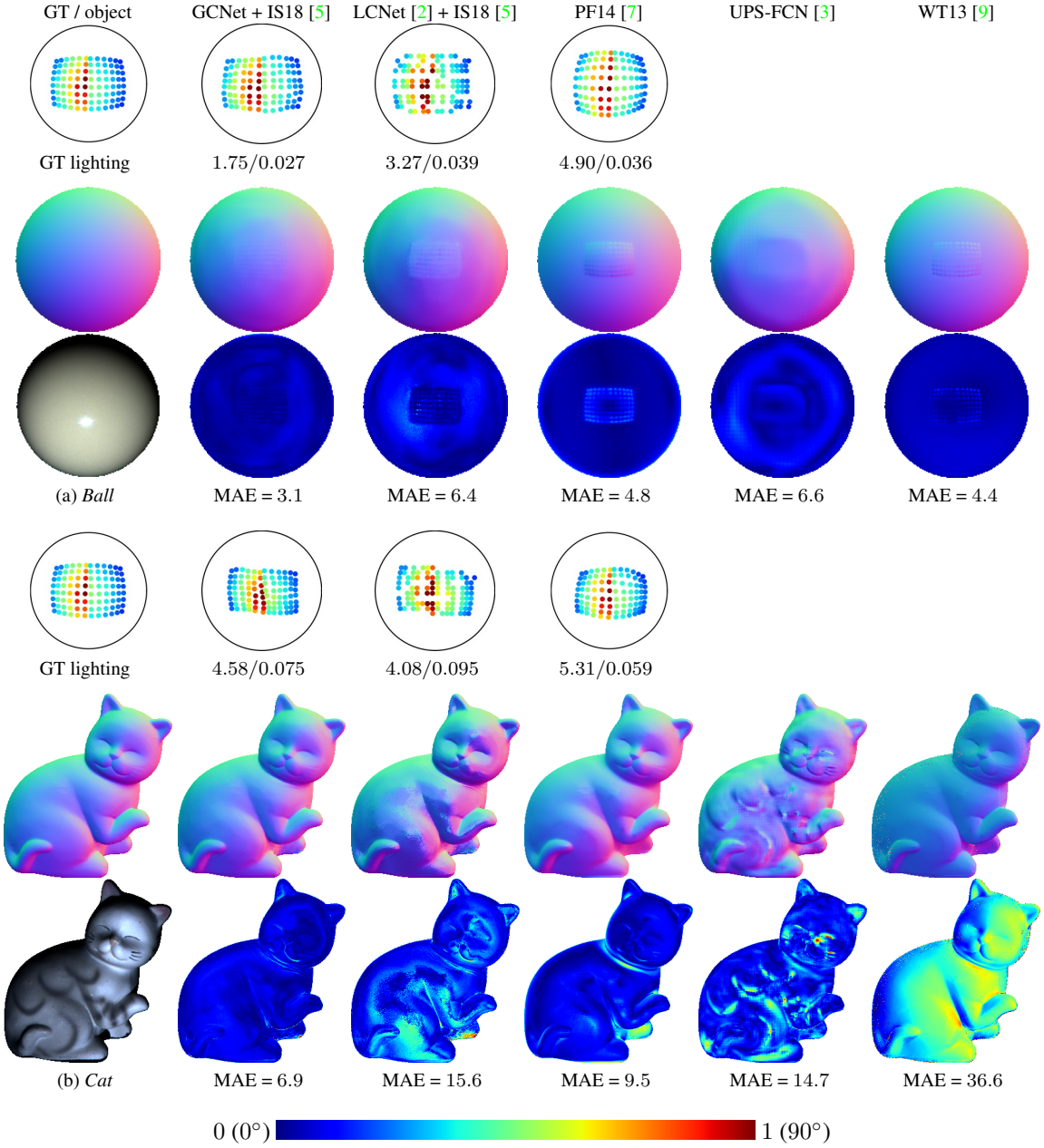


Figure S6. Results for the *Ball* and *Cat* from the DiLiGenT benchmark [8].

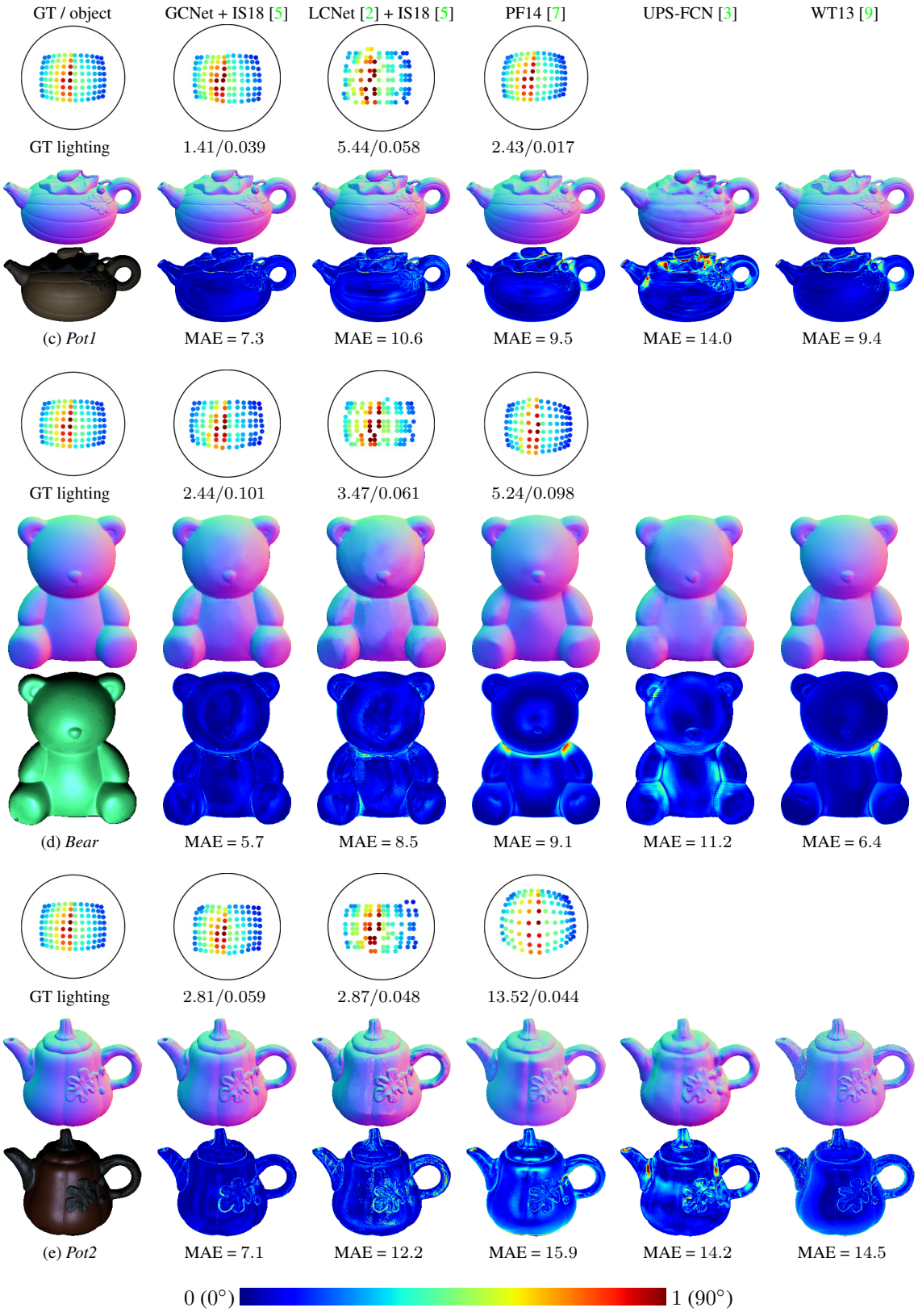


Figure S7. Results for the *Pot1*, *Bear* and *Pot2* from the DiLiGenT benchmark [8].

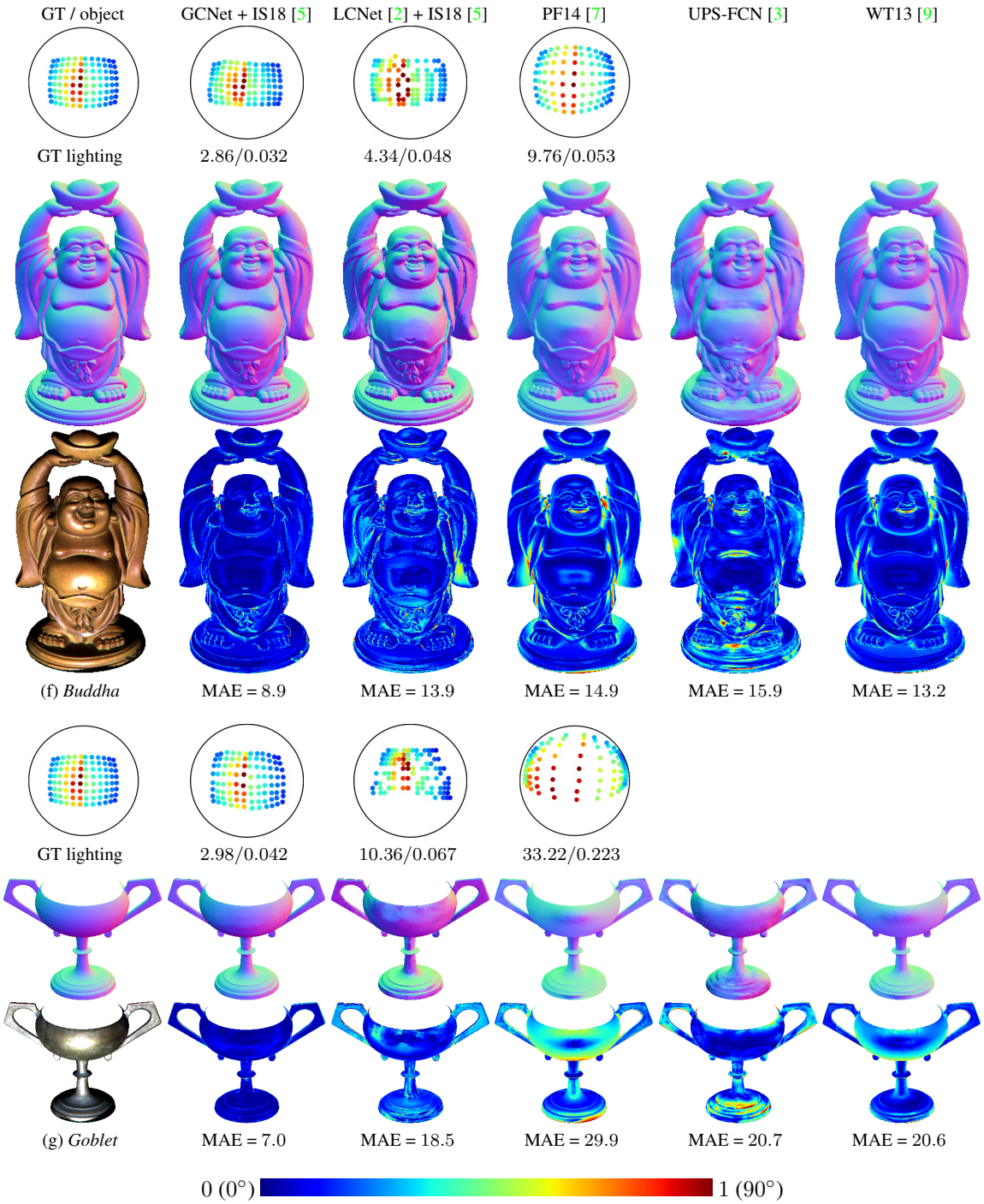


Figure S8. Results for the *Buddha* and *Goblet* from the DiLiGenT benchmark [8].

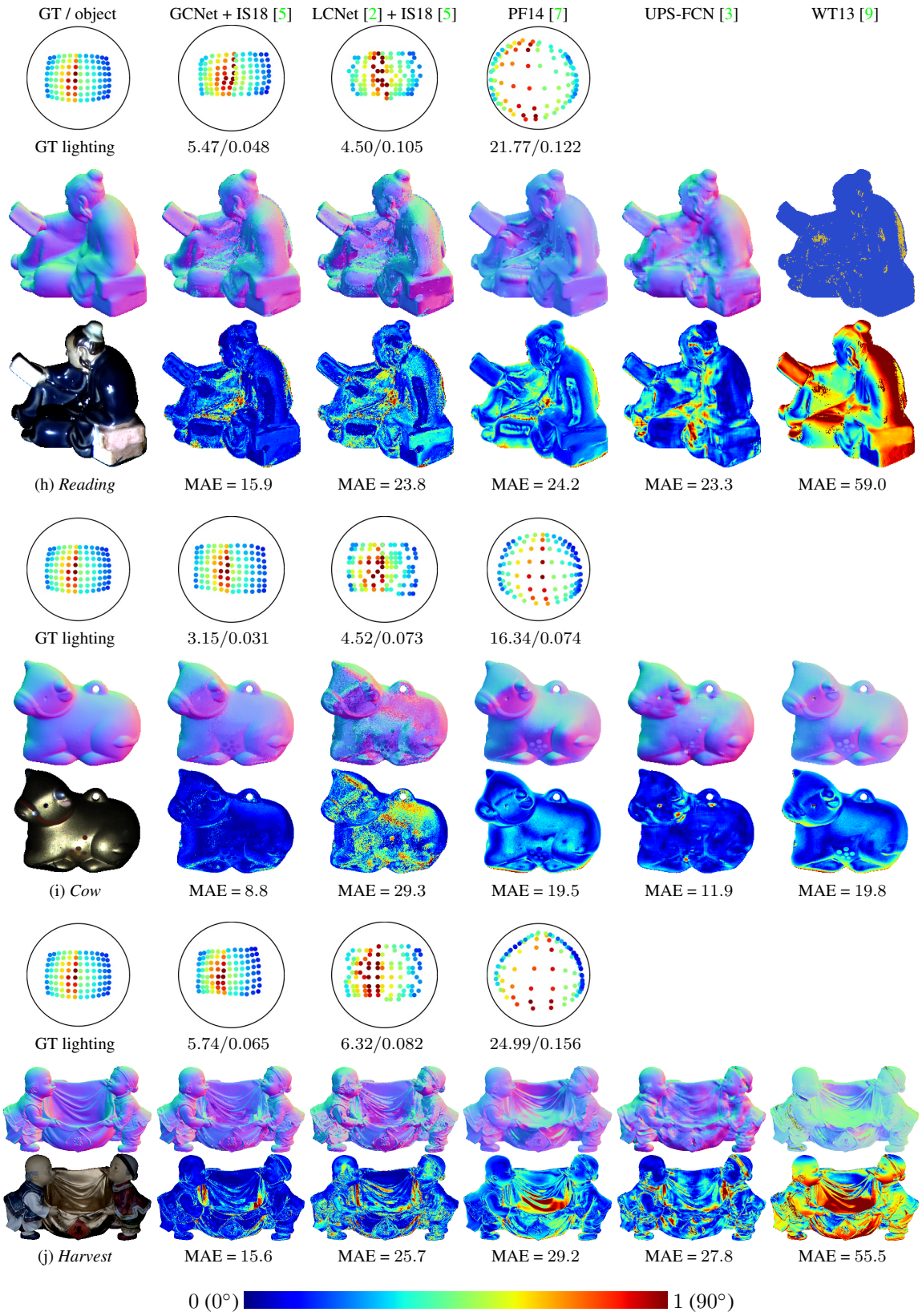


Figure S9. Results for the *Reading*, *Cow* and *Harvest* from the DiLiGenT benchmark [8].

5. Results on the DiLiGenT test dataset

We further evaluated our method on the DiLiGenT test dataset [8], which contains 9 real-world objects illuminated under 96 different light directions. These 9 objects are the same as in DiLiGenT benchmark (except that the DiLiGenT test dataset does not include the *Ball*), but captured from a different viewpoint. Table S3 and Figure S10 compare the lighting estimation results of our method with PF14 [7] and LCNet [2]. Our method achieves the best average result with an MAE of 4.92 for light directions and a relative error of 0.55 for light intensities, improving the results of LCNet by 18.0% and 42.1%, respectively.

Table S3. Lighting estimation results on the DiLiGenT test dataset. Direction and intensity are abbreviated to "dir." and "int.".

	Cat		*Pot1*		*Bear*		*Pot2*		*Buddha*		*Goblet*		*Reading*		*Cow*		*Harvest*		average	
model	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.
PF14 [7]	**4.64**	**0.086**	**2.35**	**0.020**	**3.78**	**0.028**	15.15	**0.038**	9.98	0.059	29.04	0.133	21.34	0.077	18.31	0.121	23.15	0.368	14.19	0.103
LCNet [2]	5.56	0.129	5.21	0.059	3.96	0.106	**3.15**	0.051	5.96	0.062	11.59	0.060	**4.92**	0.157	6.99	0.171	6.66	0.063	6.00	0.095
GCNet	5.93	0.088	4.51	0.049	4.15	**0.028**	3.16	0.052	**4.29**	**0.045**	**9.17**	**0.058**	5.08	**0.049**	**2.53**	**0.081**	**5.47**	**0.051**	**4.92**	**0.055**

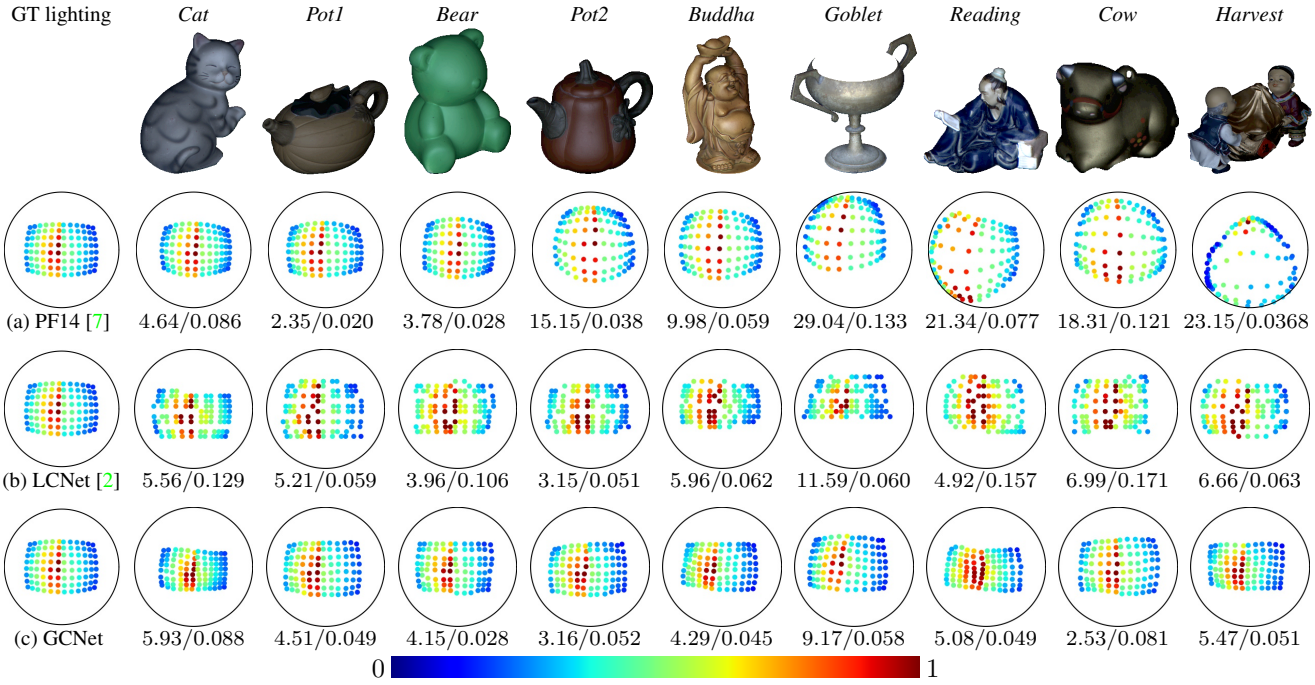


Figure S10. Visualization of the ground-truth and estimated lighting distribution for the DiLiGenT test dataset. Values below estimated lighting distributions are MAE for light direction and relative error for light intensity.

6. Results on the Light Stage Data Gallery

The Light Stage Data Gallery [4] contains 6 objects, each captured under 253 light directions of which we only used the 133 images in which the object's front was illuminated. The Light Stage Data Gallery only provides calibrated light directions and intensities, but no ground-truth normals. Figure S11, Figure S12 and Figure S13 show visual comparisons on the Light Stage Data Gallery [4].

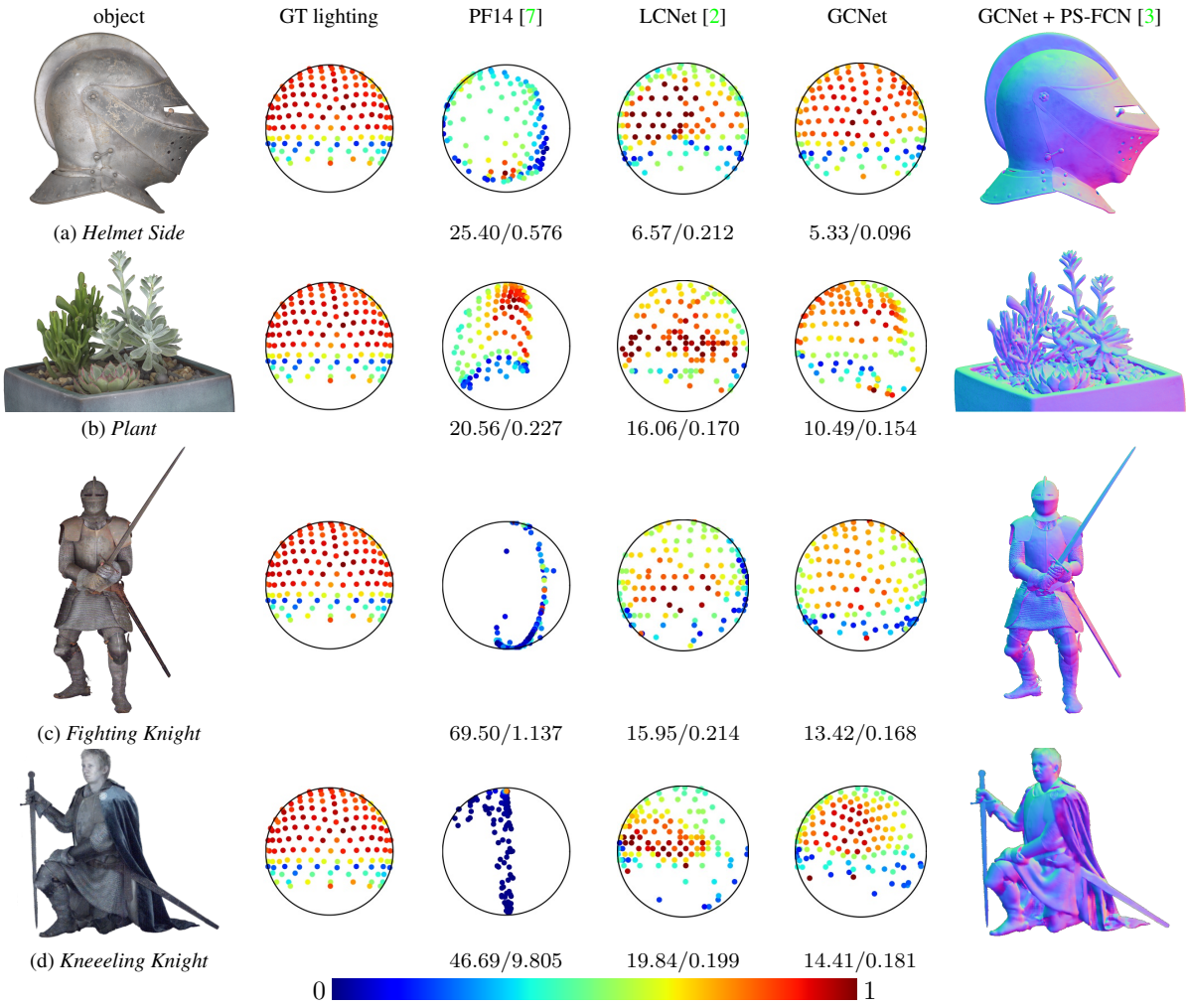


Figure S11. Results on the Light Stage Data Gallery [4]. *Column 1:* Image of the object. *Columns 2–5:* Ground-truth lightings and estimated lightings. *Column 6:* Normals predicted by PS-FCN [3] given lightings estimated by our method. The values below the estimated lightings indicate MAE for light direction and relative error for light intensity.

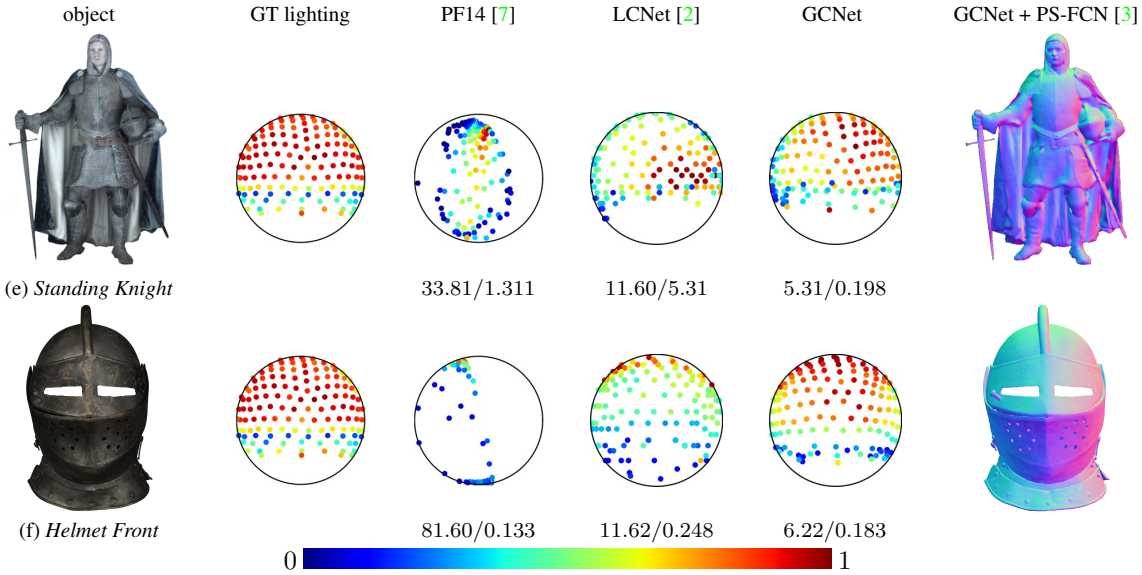


Figure S12. Results on the Light Stage Data Gallery [4]. *Column 1:* Image of the object. *Columns 2–5:* Ground-truth lightings and estimated lightings. *Column 6:* Normals predicted by PS-FCN [3] given lightings estimated by our method. The values below the estimated lightings indicate MAE for light direction and relative error for light intensity.

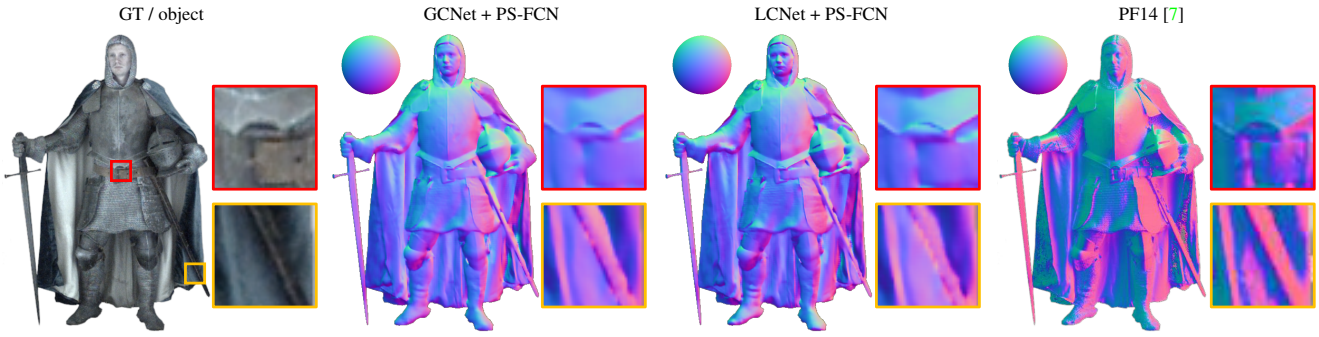


Figure S13. Visual comparison of normal estimation for the Light Stage Data Gallery's *Standing Knight*. We can see that coupled with our method's more accurate lightings, PS-FCN can produce more reliable normal estimation for thin regions.

7. Results on the Harvard photometric stereo dataset

The Harvard photometric stereo dataset [10] contains 7 Lambertian objects, each with 20 gray-scale images. LCNet [2] and GCNet have been trained with color images, as input for the networks we simply stacked the single-channel gray-scale images to form 3-channel images. Table S4 and Figure S14 show the light direction estimation results. Note that the provided images are already normalized by the calibrated intensities, and that no ground-truth normals are provided. PF14 [7] works well on Lambertian objects. Our method achieves the best average MAE in light direction estimation on this dataset.

Table S4. Light direction estimation results on Harvard photometric stereo dataset [10].

model	*Cat*	*Frog*	*Hippo*	*Lizard*	*Pig*	*Scholar*	*Turtle*	average
PF14 [7]	3.61	4.48	**4.76**	**3.32**	**3.95**	**5.15**	**3.05**	4.04
LCNet [2]	0.33	5.40	5.91	4.59	5.35	5.53	4.92	4.57
GCNet	**0.21**	**3.89**	6.07	4.26	4.49	5.50	3.34	**3.97**

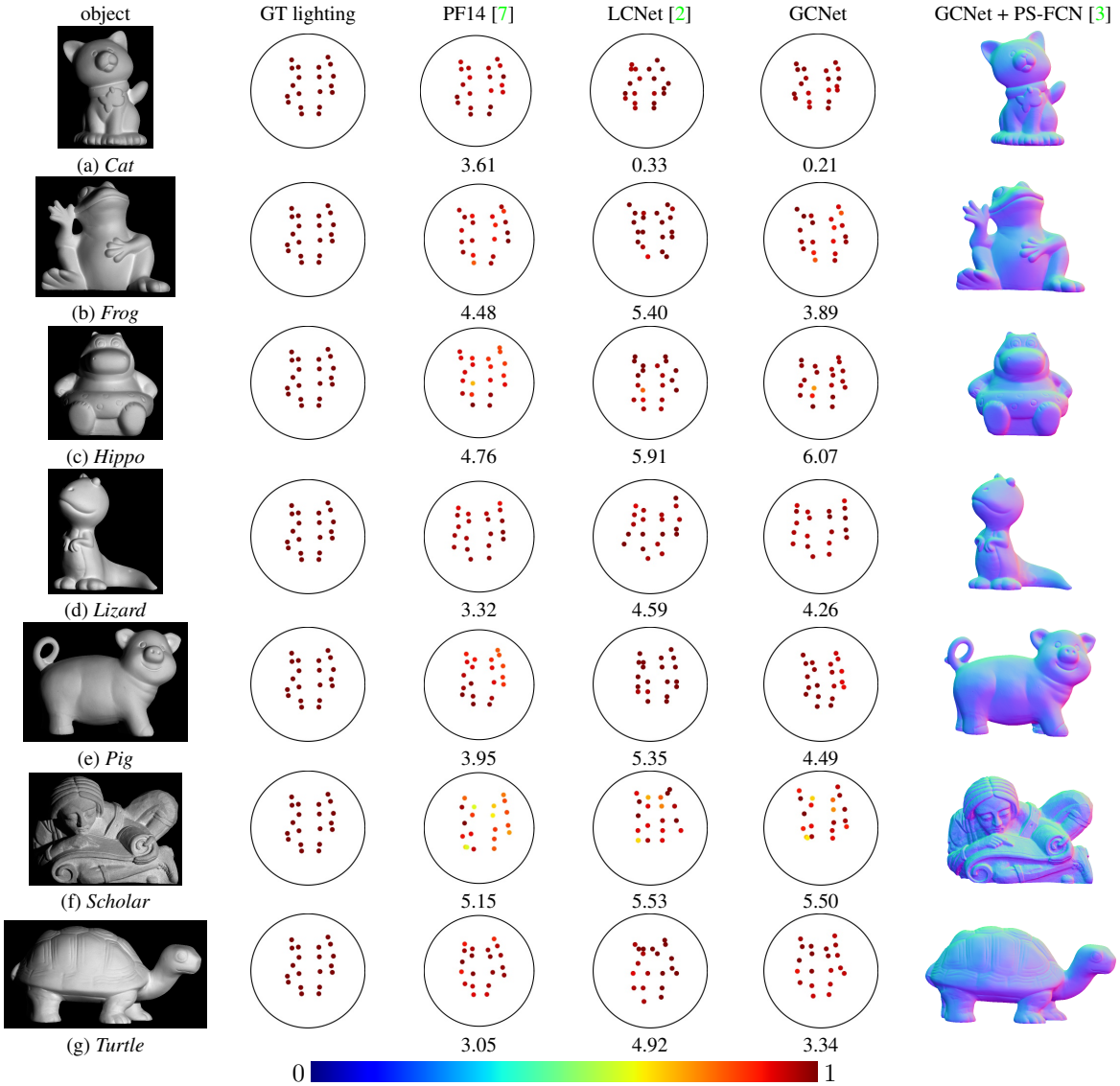


Figure S14. *Column 1:* Object. *Columns 2–5:* Ground-truth and estimated lightings. *Column 6:* Normals predicted by PS-FCN [3] given lightings estimated by our method. The values below the estimated lightings indicate MAE in light direction estimation.

8. Results on the Gourd&Apple dataset

The Gourd&Apple dataset [1] is a real dataset containing 3 objects, *Apple*, *Gourd1* and *Gourd2* with 112, 102 and 98 images, respectively. Table S5 and Figure S15 show that our method performs slightly worse than LCNet in light direction estimation but better in light intensity estimation. As there are many factors that can affect the lighting estimation results (*e.g.*, object shape, lighting distribution, surface reflectance, and noise), the exact reason why our method cannot outperform LCNet on the Gourd&Apple dataset is unclear.

However, we consider that the lighting distribution may not be the reason, because our experiment on the synthetic dataset showed that our method performs better than LCNet on different lighting distributions (both diverse and biased lighting distributions). Object shape and surface reflectance, on the other hand, may be potential reasons: The shapes in this dataset are roughly squashed spheres, and as a consequence our method works just comparable to LCNet. And since the network has been trained on a synthetic dataset rendered with 100 MERL BRDFs, it may perform worse on some real BRDFs that lie in a space that is far from the BRDF space spanned by the training data. We consider exploring more comprehensive training datasets in the future.

Table S5. Light direction estimation results on the Gourd&Apple dataset [1].

	Apple		*Gourd1*		*Gourd2*		average	
model	direction	intensity	direction	intensity	direction	intensity	direction	intensity
PF14 [7]	**6.68**	0.109	21.23	0.096	25.87	0.329	17.92	0.178
LCNet [2]	9.31	0.106	**4.07**	0.048	**7.11**	**0.186**	**6.83**	0.113
GCNet	10.91	**0.094**	4.29	**0.042**	7.13	0.199	7.44	**0.112**

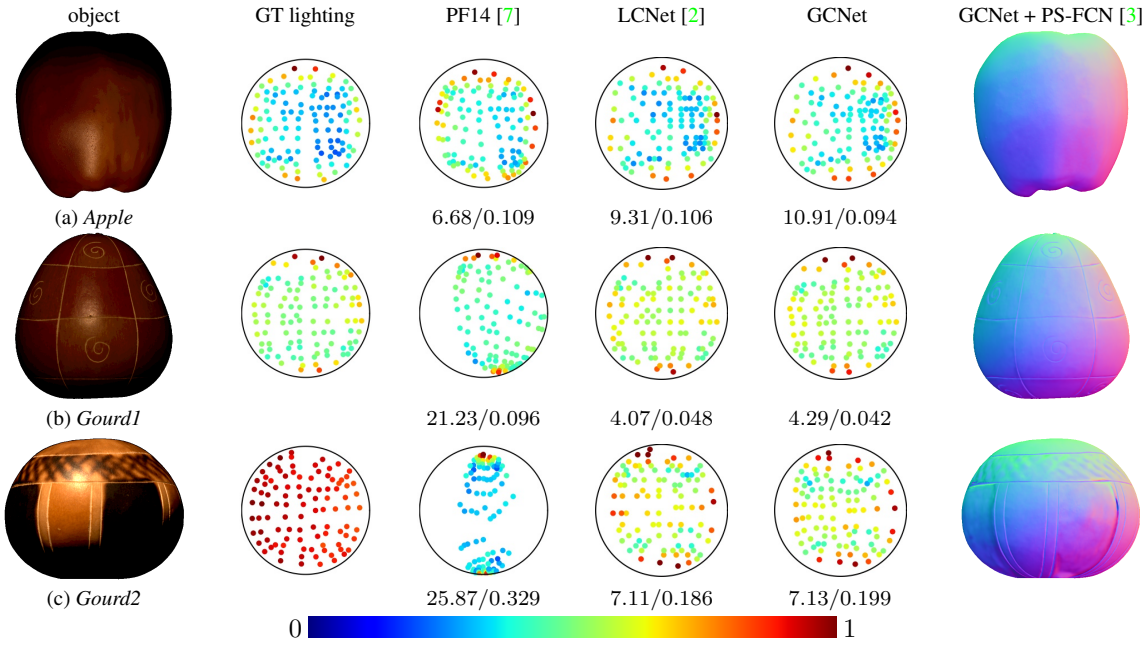


Figure S15. Results on the Gourd&Apple dataset.

References

[1] Neil G. Alldrin, Todd Zickler, and David J. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*, 2008. 1, 15

[2] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K. Wong. Self-calibrating deep photometric stereo networks. In *CVPR*, 2019. 1, 2, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

[3] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. PS-FCN: A flexible learning framework for photometric stereo. In *ECCV*, 2018. 7, 8, 9, 10, 12, 13, 14, 15

[4] Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. Relighting human locomotion with flowed reflectance fields. In *EGSR*, 2006. 1, 12, 13

[5] Satoshi Ikehata. CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In *ECCV*, 2018. 7, 8, 9, 10

[6] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. In *SIGGRAPH*, 2003. 1, 5

[7] Thoma Papadhimitri and Paolo Favaro. A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima. *IJCV*, 2014. 7, 8, 9, 10, 11, 12, 13, 14, 15

[8] Boxin Shi, Zhipeng Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo. *TPAMI*, 2019. 1, 7, 8, 9, 10, 11

[9] Zhe Wu and Ping Tan. Calibrating photometric stereo by holistic reflectance symmetry analysis. In *CVPR*, 2013. 7, 8, 9, 10

[10] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J. Gortler, David W. Jacobs, and Todd Zickler. From shading to local shape. *TPAMI*, 2014. 1, 14