Supplementary Material for "MDCN-PS: Monocular Depth estimation aware Coarse Normal attention for Robust Photometric Stereo"

Masahiro Yamaguchi, Takashi Shibata, Shoji Yachida, Keiko Yokoyama and Toshinori Hosoi
Visual Intelligence Research Laboratories, NEC Corporation
1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa 211-8666, Japan
masahiro-yamaguchi@nec.com, t.shibata@ieee.org, s-yachida@nec.com
k.yokoyama@nec.com, t.hosoi@nec.com

This is the supplementary material for "MDCN-PS: Monocular-Depth-guided Coarse Normal attention for Robust Photometric Stereo" (our main paper). We provide details of the datasets, additional experimental results, implementation details, limitations, and feature works.

A. Datasets

A.1. Example of PS Objaverse Dataset

We present an example of the *PS-Objaverse* dataset (Sec. 3.3 in main paper). The *PS-Objaverse* dataset consists of ten RGB images with a resolution of 512×512 and the corresponding ground truth normal images, comprising a total of 30,488 scenes. Figure 1 shows the ten RGB images from the *PS-Objaverse* dataset and their corresponding ground truth normals.

A.2. Detail of Blender Settings

We generated a novel dataset, the *PS-Objaverse* dataset, utilizing 3D models from the Objaverse dataset. Figure 2 shows the flow of *PS-Objaverse* dataset construction. This paper aims to create a high-quality dataset through photorealistic rendering using Blender's Cycles rendering engine [2] (Sec. 3.3 in main paper). We randomly place ten point light sources or directional light sources on the upper hemisphere. Then, we render an RGB image with a resolution of 512×512 pixels for each light source. One to five objects are randomly positioned in each scene, and a randomly selected material is assigned to each object. Here, the materials are randomly chosen from the ambientCG [1]. Note that, unlike IS23 [8], it is unnecessary to guarantee that material categories do not overlap within the same scene. We use the Principled BSDF to read color, roughness, and metallicity maps from each texture file in the material setting. Note that, normal and displacement maps are not used in this setting, as their influence at a resolution of 512×512 is too detailed, resulting in artifacts. A composite tree is configured for each scene to create ground truth normal maps and depth maps. The normal maps are used for evaluation during training and testing, while the depth maps are used to measure the effective number of light sources during testing. Other detailed settings are summarized in Tab. 1.

Table 1. Detail of Blender Parameters

Parameter	Value
camera.type	ORTHO
exposure	0
gamma	1.0
samples	256
max_bounces	10
diffuse_bounces	10
glossy_bounces	10
transmission_bounces	10
volume_bounces	10
use_denoising	True

B. Additional Results

B.1. Additional Results on Synthetic Dataset

In the section "Results on Synthetic Data using CGTrader [3] and Share Texture [4]" (Sec. 4.2 in our main paper), ten objects expected to cast shadows were selected from CGTrader. For each object, a texture was carefully chosen from ShareTextures and assigned into the following six categories: Plaster, Wood, Metal, Floor, Fabric, and Plastic. The three light sources were positioned as viewed from directly above, as shown in Fig. 3 (a), and the camera was placed at the center of the circle.

We show the additional results corresponding to the "Results on Synthetic Data using CGTrader [3] and Share Tex-

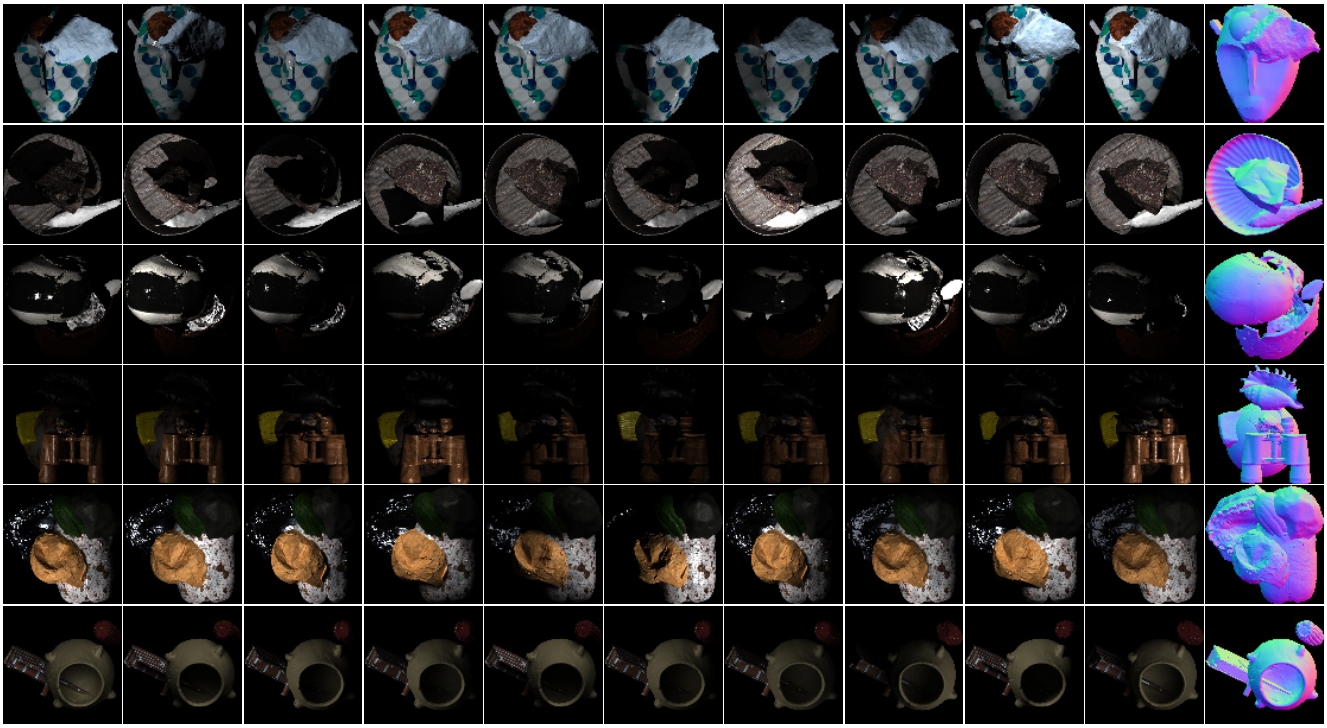

Figure 1. Examples of images and normals in our *PS-Objaverse* dataset

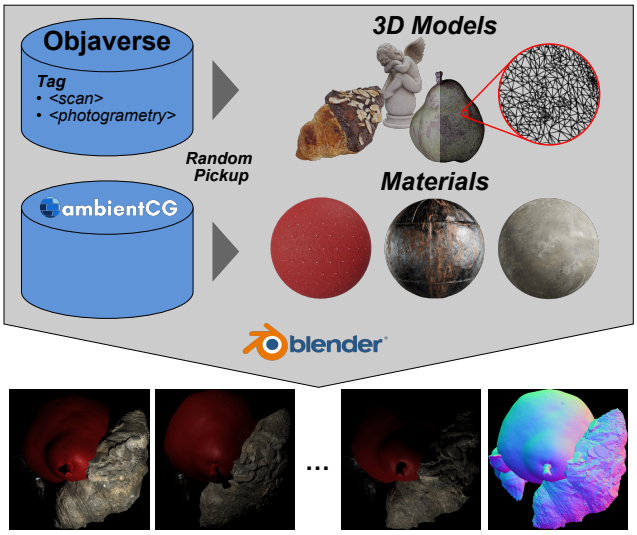

Figure 2. Workflow for the *PS-Objaverse* dataset construction. 3D models selected from the Objaverse [7] and materials randomly picked from ambientCG [1] are rendered using Blender [2]. During the selection of 3D models, models scanned from real-world data are filtered by tags, and models with a higher number of vertices compared to 3D models are prioritized. The lower part shows the generated images with their corresponding normal maps.

ture [4]" (Sec. 4.1 in our main paper). Figure 4, and 5, similar to Fig. 5 in our main paper, shows the input images, the

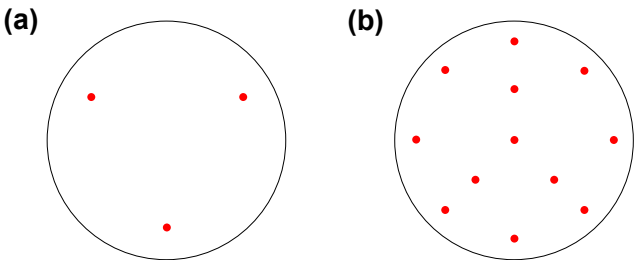

Figure 3. Light location: (a) Synthetic Dataset, (b) Photogrammetry Dataset.

lighting conditions for each region, the results of the proposed method, the results from IS23 [8], and the ground truth normal information, respectively. As mentioned in the main paper, it can be confirmed from Fig. 4, and 5 that our method achieves accurate normal estimation, even in areas where the light source is insufficient. These results show that the proposed method is effective even under challenging lighting conditions.

B.2. Additional Results on the DiLiGenT dataset

We present a qualitative evaluation using two input images from the DiLiGenT dataset [9] (Sec. 4.2 in our main paper). The estimation results of HARVEST and READING are shown in Fig. 6. These objects, known for their highly non-convex geometry and tendency to create

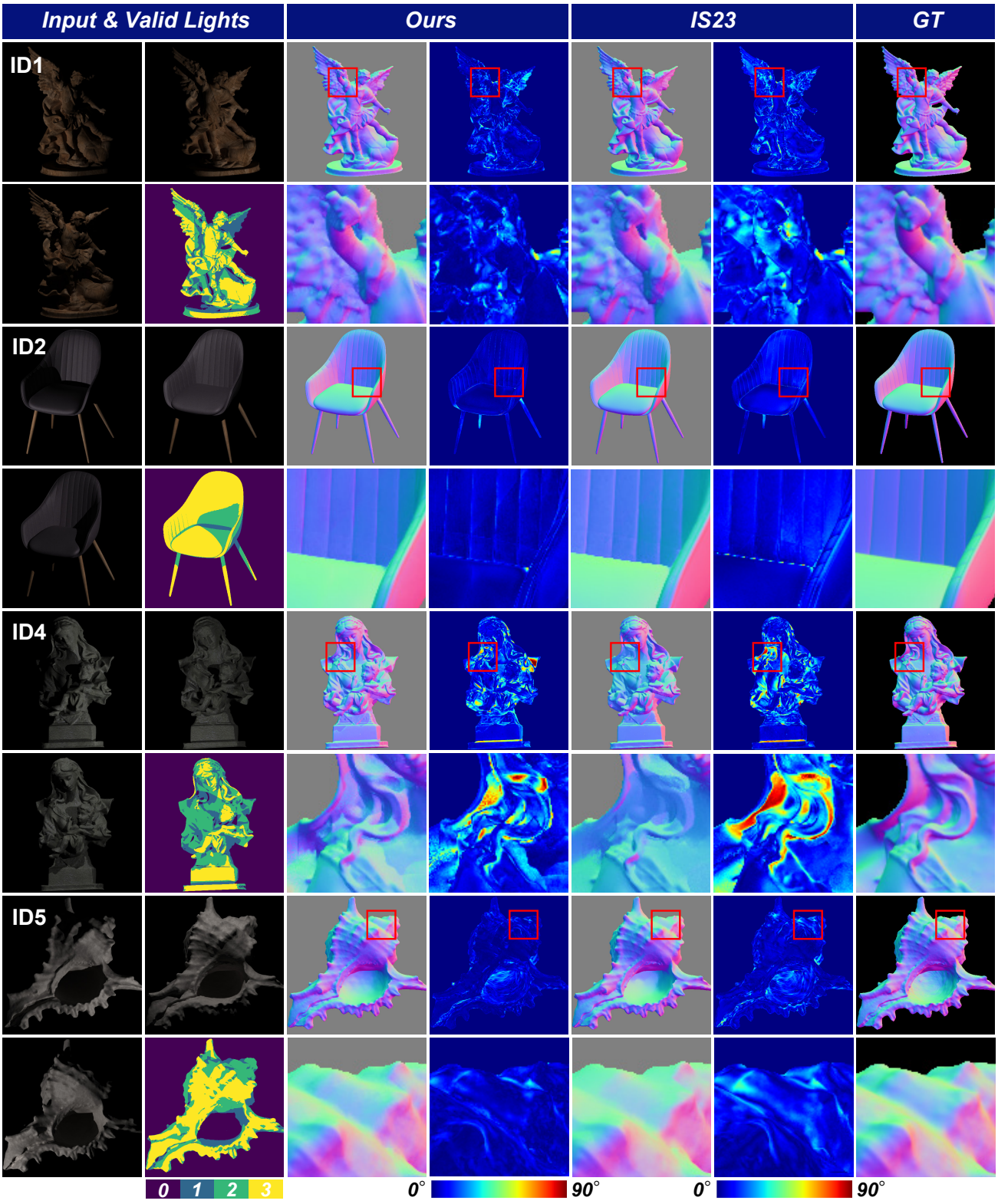

Figure 4. Qualitative evaluation of our synthetic dataset for ID1, ID2, ID4, ID5. The input images, the number of effectively illuminated light sources for each pixel, the estimation results by our method, the estimation results by IS23, and the ground truth are shown, respectively. Each estimation result indicates the error between the estimated normals and the ground truth.

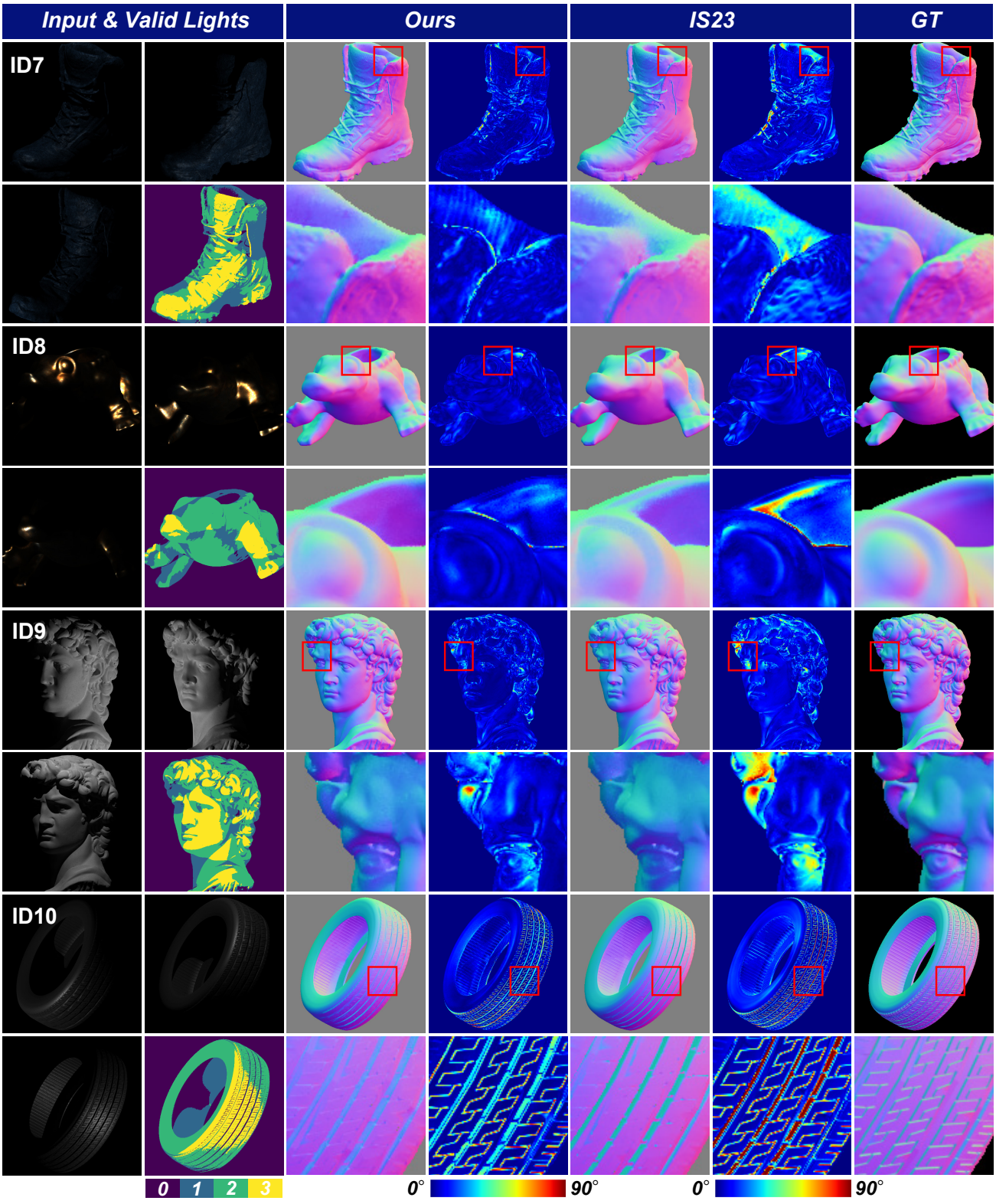

Figure 5. Qualitative evaluation of our synthetic dataset for ID7, ID8, ID9, ID10. The input images, the number of effectively illuminated light sources for each pixel, the estimation results by our method, the estimation results by IS23, and the ground truth are shown, respectively. Each estimation result indicates the error between the estimated normals and the ground truth.

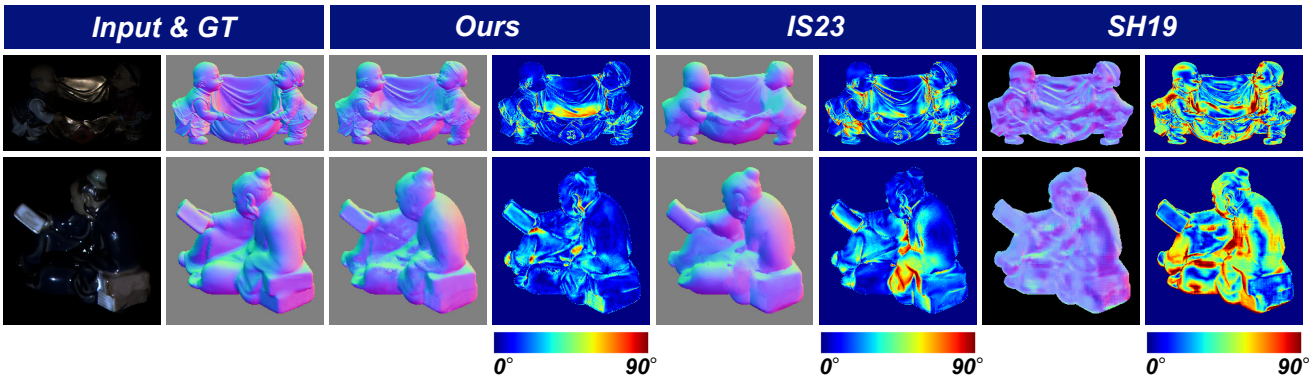

Figure 6. Qualitative evaluation on the DiLiGenT dataset. The input images, the estimation results by our method, the estimation results by IS23, the estimation results by SH19, and the ground truth are shown, respectively. Each estimation result indicates the error between the estimated normals and the ground truth.

shadows, are considered the most challenging benchmarks, adding significant difficulty to the estimation. Figure 6 shows the results of the proposed method, the results from IS23 [8], the results from CH19 [6], and the ground truth normal information, respectively. These results also show the superiority of the proposed method over the conventional methods in visual comparisons. Note that, for each scene, 001.png and 002.png were used as input images.

B.3. Additional Results on Photogrammetry dataset

In the evaluation without object masks, 3D reconstruction was performed using photogrammetry on real image data in the Mip-Nerf360 dataset [5]. The 3D data was then rendered using Blender. When rendering, the texture from the 3D reconstruction is applied to the Base Color of the Principled BSDF. Figure 3 (b) shows the arrangement of the light sources as viewed from above, with the camera positioned at the center of the circle. During rendering, the camera position was adjusted so that the model occupies the entire field of view. Special attention was paid to ensuring that the foreground and background were distinguishable.

We present the remaining results, i.e., counter and stump, corresponding to the "Evaluation without Object Mask" (Sec. 4.4 in our main paper). Figure 7, similar to Fig. 6 in the main paper, shows examples of input images, the results of the proposed method, the results from IS23 [8], and the ground truth, respectively. As mentioned in the main paper, it can be seen from Fig. 7 that our method has improved normal estimation accuracy, particularly in regions classified as background. These results show that our method is effective even in scenarios involving background regions.

C. Implementation

C.1. Implementation Details

We describe the implementation details of the proposed method. The training loss was calculated using MSE loss, where the $l2$ error between the predicted and ground truth surface normal vectors was computed. The accuracy was evaluated based on the mean angular error (MAE) between the predicted normal map and the ground truth normal map, with angles measured in degrees. The batch size was set to 8, with an initial learning rate of 0.0001 and a weight decay of 0.05. The number of input training images per batch was randomly selected between 3 and 6. We use fixed weights for Depth Anything V2 and use a pre-trained model named "Depth-Anything-V2-Large". The training was performed on four NVIDIA V100 cards for approximately seven days. The inference time depends on the number and resolution of the input images.

The dataset is augmented during training to introduce more variation in the training examples. Specifically, we randomly flip the images horizontally or vertically or rotate them by 90 degrees. All data augmentations are applied with a 50% probability.

C.2. Limitations and Future Works

Finally, we discuss the limitations and future works of the proposed method. Monocular Depth Estimation is not always a panacea; generally, a depth obtained by monocular depth estimation can be erroneously estimated in confusing targets, such as photorealistic paintings. The monocular depth may not be a practical guide for Coarse Normal attention for such targets.

Monocular depth estimation is typically trained using a dataset based on the perspective view. Note that while the obtained coarse depths may contain errors due to the projection method, the proposed method, which uses the coarse

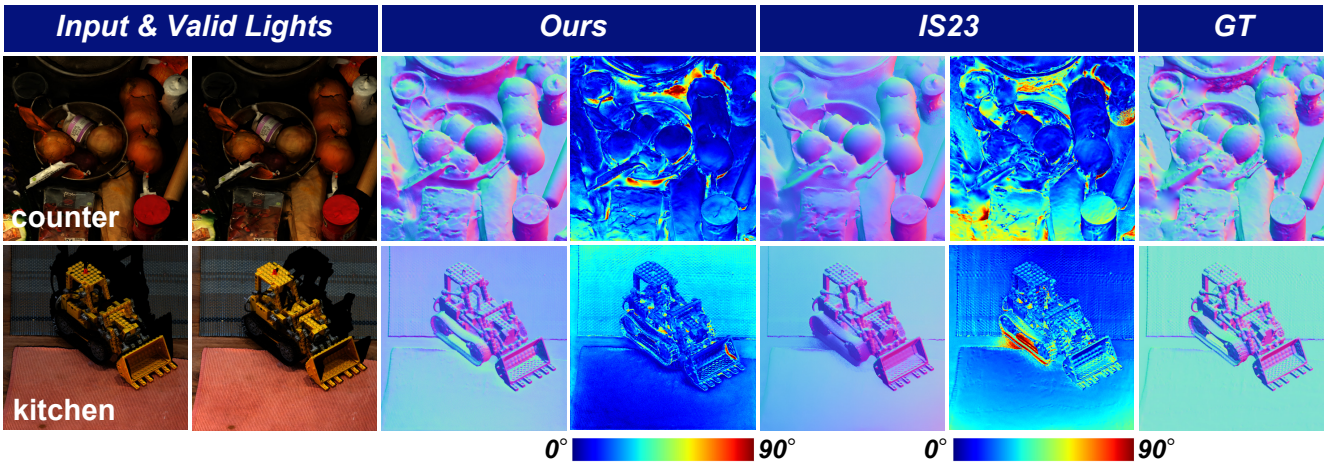

Figure 7. Qualitative evaluation of a dataset using the 3D model reconstructed by our photogrammetry. The input images, our method's estimation results, IS23's estimation results, and the ground truth are shown. Each estimation result displays the error between the estimated normals and the ground truth.

normal as a guide in Coarse Nomal attention, outperforms the state-of-the-art existing methods [8] as described in our main paper. This fact suggests that, in the future, the performance of the proposed method could be further improved by developing monocular depth estimation to reduce errors due to the projection protocol.

References

[1] ambientcg. https://ambientcg.com/. 1, 2

[2] Blender. https://www.blender.org/. 1, 2

[3] Cgtrader. https://www.cgtrader.com/. 1

[4] sharetextures. https://www.sharetextures.com/. 1, 2

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5470–5479, 2022. 5

[6] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8739–8747, 2019. 5

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 13142–13153, 2023. 2

[8] Satoshi Ikehata. Scalable, detailed and mask-free universal photometric stereo. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 13198–13207, 2023. 1, 2, 5, 6

[9] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3707–3716, 2016. 2