Efficient light transport acquisition by coded illumination and robust photometric stereo by dual photography using deep neural network

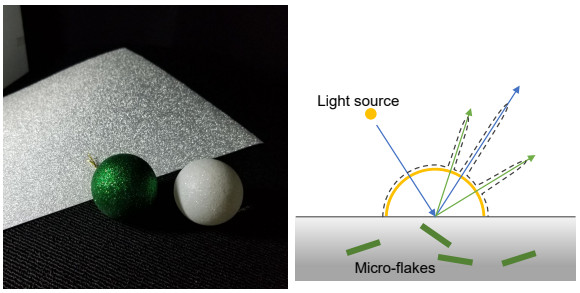
Takafumi Iwaguchi and Hiroshi Kawasaki
Kyushu University
Fukuoka, Japan
iwaguchi@ait.kyushu-u.ac.jp, kawasaki@ait.kyushu-u.ac.jp

Abstract

We present an efficient and robust photometric stereo (PS) measurement by a setup with an optical diffuser. The setup which includes a single projector that places point light sources on the diffuser, extends the possibility of flexible measurement without the limitation from employing physical light sources, i.e. *a number and illumination shape. By taking advantage of the setup, we design an illumination for effective and robust surface normal estimation. Unlike the previous techniques, we utilize deep neural network consists of a renderer and a PS module to find multiplexed illumination patterns, which are suitable for PS measurement. Another challenging problem is to measure objects with micro-structures which reflect the light randomly according to lighting and viewing directions. To overcome the problem, we propose a novel PS measurement using a dual-photography setup, which allows us to analyze the angular distribution of reflection by capturing reflection pattern on the diffuser. We show a smooth surface normal can be estimated by simply applying a low-pass filter on the captured images. Moreover, we also propose an effective sampling to deal with time-consuming measurement of dual photography setup. We show that by utilization of the trained sampling codes by DNN considering light transport in the setup, the number of the measurement is drastically reduced.*

1. Introduction

Light transport (LT) is fundamental information to describe both photometric and geometric information of a scene, however, to obtain the full LT, it is required to capture the entire scene by turning on every single pixel of the video projector one by one. It is problematic not only with time consuming for data acquisition, but also extremely low SNR, since only single pixel is projected at each capture. Hadamard coding has been proposed for solving SNR problem [20], however, the same number of measurements is still needed. A compressive sensing has been proposed to reduce the measurement time [24, 23], however, still a large number is required to obtain a sufficiently accurate LT. In this paper, we propose a method to drastically reduce the number of measurements by designing the projecting pattern specialized for PS using DNN. To efficiently capture the surface reflectance, we use a single fixed video projector as the light source, a single fixed camera and an optical diffuser which is placed between the camera and the target object. A two-dimensional distribution of intensity on the diffuser plane in the captured image can be interpret as an angular distribution of the reflected light.



(a) (b)

Figure 1. (a) Glitter paper exhibits unpredictable strong reflection due to micro-flakes locate randomly inside. (b) Illustration of specular robe.

By using LT, the scene lit by an arbitrary lighting condition can be synthesized, which will be utilized for various purposes. Among them, photometric stereo (PS) is one important application, which requires multiple images captured under different lighting positions. Common photometric stereo techniques assume Lambert model on object surfaces and infinitely distant light sources. The surface normal is efficiently recovered by a linear system if images are captured by a camera with at least three different light source positions. For non-Lambertian object which can be approximated by a simple model, such as specular reflection, a non-linear method can be applied [11], however, for more complex BRDF, it is rarely solved by PS. One such material is a glitter paper, which has a complex micro-

structure on its surface, as shown in Fig. 1(a). A typical glitter paper consists of micro-flakes with strong reflection, which are distributed randomly, resulting in unpredictable anisotropic reflection with respect to the light and view direction. Especially, when a single pixel of a captured image contains more than one reflection probes towards various directions, it cannot be assumed as diffuse surface (Fig. 1(b)) and the luminance of neighboring pixels changes almost independently of each other. In the paper, we exploit this angular distribution to extract reflections that contributes to surface estimation by using dual photography, which can swap the role of the camera and the projector and to synthesize *dual views*, by using LT. Since the lighting of the dual views is defined by the area in the captured image, the scene can be relit with arbitrary illumination.

Contribution of our paper are summarized as follows:

- The number of required projection pattern to acquire LT can be drastically reduced by generating patterns optimized by DNN.
- PS can be applied to a single fixed video projector by using dual photography technique and a optical diffuser in front of the camera.
- PS can be applied to the surface with complicated BRDF, such as glitter papers, by applying low-pass filter on the reflected light received on the diffuser as a post-process so that specular reflections from micro-flakes are removed.

2. Related work

2.1. Light transport and dual photography

Sen *et al*. proposed dual photography (DP), an imaging technique to interchange light and camera in a scene [21]. The authors demonstrated that relighting using any pixel of an image captured by a camera as a light source. For relighting, LT matrix between each pixel of the projector and each pixel of the camera is sampled. One of the difficulties with DP is that the LT is huge due to the resolution of the camera and projector. When the most of the objects in a scene have a diffuse surface, LT is sparse, *i.e*., light ray starts from a light source falls in a single pixel of a camera. In this case, hierarchical sampling has been proposed in the original paper and a sampling method using compressive sensing has also been proposed [22, 17] for efficient sampling. Since the LT are dense in our setup with diffuser, our goal is to efficiently sample the dense LT.

In recent years, by combining high-speed digital spatial light modulators with various sensors, DP has been applied to new imaging methods. For example, DP has been used in photography using photodetectors [1, 6], hyper spectral imaging using spectrometers [7], and depth measurement using time-resolved sensors [25]. These are special cases of DP where the camera has a single pixel.

There have been studies on PS with multiple photodetectors [24, 30], and on PS when the diffuser is between the object and the photodetectors [23]. While these methods are in principle the same as measurements using a small number of isolated light sources, we achieve a dense arrangement of many light sources, which makes it possible to filter the light rays as shown in the experiments.

2.2. Light multiplexing and optimal coding for efficient measurement

Light multiplexing is a common technique in active measurement to improve energy efficiency. To improve SNR when the energy of the light source is small, as in the case of LT measurements, multiplexing using Hadamard code has been shown to be optimal for the number of the measurement [20]. In the paper, the authors projected Hadamard code to diffuse wall and successfully performed PS using the weak reflected light, by taking a full advantage of high SNR. On the other hand, we directly use optical diffuser which is advantageous on energy efficiency compared to reflection by wall; note that it greatly helps to decrease the number of projected patterns, which is optimized by DNN.

In terms of decreasing the number of measurement, compressed sensing has been intensively researched, such as for triangulation purpose [5], and imaging techniques based on primal dual coding where illumination and exposure are coded simultaneously [16, 15]. Kang et al. [10] used multiplexing to reduce the number of measurements when acquiring BRDF using a device with a high-density LED array. DNN is utilized to learn how to efficiently multiplex light and decode BRDFs from observed images when the number of measurements is limited. Although the concept of measuring LT by optimizing lighting pattern by DNN is similar to [10], since the measured image through the diffuser is blurred, new technique is required. In the paper, we propose new network architecture as well as binary codes to increase the difference in observations under blur.

2.3. Photometric stereo for non-Lambertian surfaces

One of major problems on PS is reconstruction of non-Lambertian surfaces. Specular reflections have been dealt with by decomposing reflection components [13], by using more than four light directions [2], or by applying a median filter [12]. PS for materials represented by a BRDF model has also been proposed [11, 4]. Since a direction of reflection of glitter materials changes randomly depending on the direction of the flakes, it is difficult to represent it by using a general BRDF model.

In recent years, learning-based PS algorithms have been proposed [29, 28, 19, 3]. In particular, DNN-based methods

can be trained on synthetic data to perform PS on a variety of materials with data augmentation. However, as the number of parameters for material increases, the number of data required to achieve sufficient accuracy increases rapidly. In addition, since synthesis of glitter surfaces with sharp specularity requires long rendering time, we propose a technique to directly apply filtering on light rays to mitigate the effect of glitter surfaces for PS in the paper; note that filtering of light rays instead of RGB image is usually difficult and it is not obvious. Using a light filed camera is one solution for such purpose [27] and PS is conducted for glossy objects [26] or under ambient lighting condition [14]. In our method, we use a common camera as well as inexpensive diffuser and apply dual photography to apply filtering to light rays.

3. Overview of the technique

3.1. System configuration

We use the setup consists of a camera, projector and optical diffuser as shown in Fig. 2. When the diffuser can be assumed an ideal diffuse transmitter which takes in light from all direction, and emit light equally to all direction of opposite side, a point where radiated by impulse beam from the projector can be considered as a point light source. We call such a setup "primal" as shown in Fig. 2(a) and it can be applied to PS measurement by simply projecting impulse beam to anywhere on the diffuser and by capturing shading images for each illumination. A surface normal is estimated by any reconstruction algorithm considering a light direction from each point of radiation.

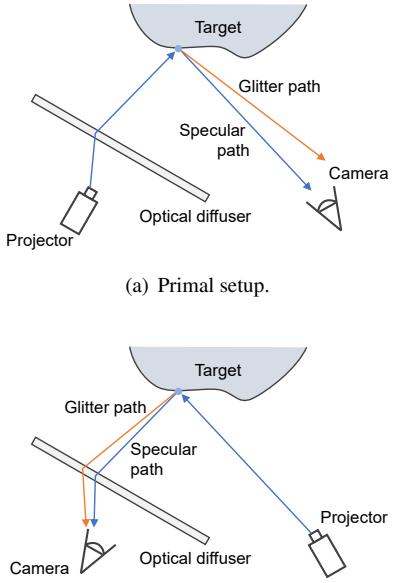
For dual photography, as shown in Fig. 2(b), this setup consists of the same components as primal setup, however, a camera and projector are swapped. A target object is directly illuminated by the projector, reflected light is cast on the diffuser, and it is captured by the camera. As described in [21], scene light transport of primal setup is written as

$$\mathbf{c}' = \mathbf{T}\mathbf{l}', \tag{1}$$

where $\mathbf{c}'$ denotes the image captured by the camera, $\mathbf{l}'$ denotes the projected pattern, and $\mathbf{T}$ denotes the light transport matrix. Observations from the viewpoint of the projector are generated according to Helmholtz reciprocity:

$$\mathbf{l}'' = \mathbf{T}^{\intercal}\mathbf{c}'', \tag{2}$$

where $\mathbf{l}''$ denotes reconstructed observation from the viewpoint of the projector, and $\mathbf{c}''$ denotes a projection pattern from the camera. Here, we can relight the scene with virtual projection pattern $\mathbf{c}'$. If we turn on a single pixel of $\mathbf{c}'$ in our setup, we can observe a scene relit with a virtual point light on the diffuser. The PS measurement can be performed for both primal and dual setup by putting a virtual point light source at an arbitrary position after the measurement. One advantage of the dual setup is that a typical camera has a wider FOV than a projector, and each virtual light source can be placed father away from others.



(a) Primal setup.

(b) Dual setup.

Figure 2. Primal and dual setups.

3.2. Algorithm

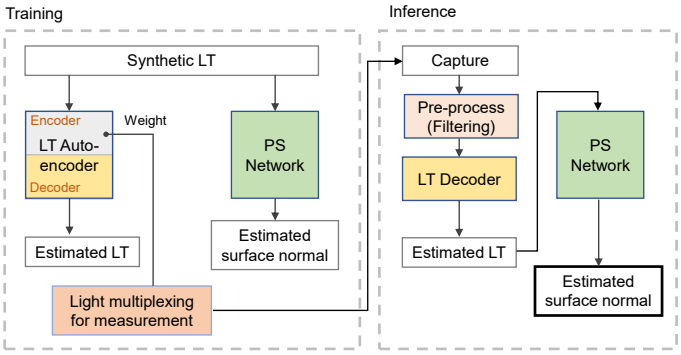


Figure 3. An algorithm overview.

An overview of our algorithm is shown in Fig. 3. In the training, since preparation of a large amount of real data of objects and their surface normal information with light source calibration data is not easy, we synthesize the scene using computer graphics and use synthetic data for training. To learn light multiplexing and decoding of LT from the measurement, we train *LT auto-encoder* by using *light multiplexing* information as a weight of network as shown in the left-hand side of the figure (detail will be explained in Sec. 3.3). *PS Network* is also trained with the synthetic

LT with ground-truth surface normal in parallel (detail will be explained in Sec. 3.4).

In the measurement, the scene is projected by special patterns, which are the weights of the encoder in the training process, using a video projector and captured by a camera. If a post-process for light rays is required, such as smoothing lights, it can be applied here. Then, shading images are reconstructed using *LT decoder*. Note that, in our method, instead of reconstructing the entire LT, each shading image under arbitrary lighting condition and view position is reconstructed in our method. The decoder network is pre-trained by synthetic data followed by fine-tuning using real data to compensate global illumination or other complicated effects. Finally the surface normal is estimated by using *PS Network*.

3.3. Learning light multiplexing for LT measurement

To improve SNR of LT measurement and to reduce number of captures, we employ light multiplexing. Typical measurement with light multiplexing is illustrated in Fig. 4. Multiple patterns are projected onto the scene and corresponding images are captured by the camera. When the scene is observed through an optical diffuser, it integrates the diffuse reflections from all the directions. Therefore, the captured image is usually blurred, and it results in dense LT; it is a major difference from LT in the scene without diffuser where the LT matrix is almost sparse.

We design light multiplexing to measure LT, or to visualize the scene behind the diffuser by capturing reflections on the diffuser. The reconstruction of full LT requires the same number as the light sources, and it can be reduced with compressive sensing techniques, however, it is not applicable for dense LT, such as our case. Instead, we find efficient multiplexing by learning a relationship among a scene, lighting, and measured intensity on the diffuser. A key idea here is that the capture can be regarded as a physical encoding, *i.e.* measured intensity on the diffuser is an encoded feature of the reflectance under a specific lighting, and the reconstruction of the scene from the captured images can be regarded as a computational decoding. We train a *LT auto-encoder* shown in Fig. 3, which combines the physical encoder and a computational decoder so that the light multiplexing and decoding are learnt jointly. The decoder is implemented using DNN as described in Sec. 4.1.

3.4. Learning PS reconstruction with dual photography setup

For PS reconstruction, we use DNN in the paper. Since we can generate all the shading images under different lighting conditions, the surface normal can also be estimated by using PS solver based on linear algebra, however, DNN-based method still has several important advantages. First,

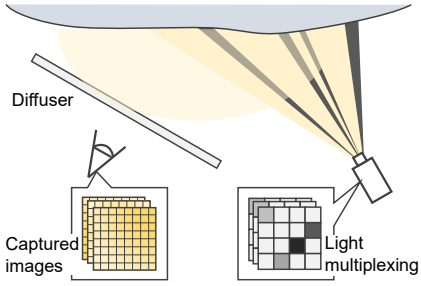


Figure 4. LT measurement with the multiplexed lighting.

since the diffuser is placed close to the object, the light source cannot be regarded as infinite, and it is difficult to estimate the correct normal by linear method. Second, we can also learn normal estimation consistently with the LT estimation by using the same geometry and setup when LT auto-encoder and light multiplexing are learnt.

Note that we also implemented end-to-end network where LT auto-encoder and PS Network are combined to learn the light multiplexing and direct PS reconstruction from the measurement simultaneously. The network is used for comparison in the experiment to verify that our separate network using each loss is better than such simple end-to-end network.

4. Implementation

4.1. Efficient measurement of LT for dual photography

The proposed DNN structure for LT measurement is shown in Fig. 5. The encoding part of LT auto-encoder can be considered as generating a view from the camera when the scene is relit by the projector using light multiplexing. Denoting patterns for light multiplexing patterns p_i, Eq. (1) is rewritten as

$$\mathbf{c_i} = \mathbf{T}\left(\mathbf{p_i} \circ \mathbf{1}\right), \tag{3}$$

where $\circ$ represents Hadamard product and $\mathbf{1}$ is a vector of 1, whose length is the number of projector pixels. Therefore, encoding layer is implemented as a matrix product with $\mathbf{p_i}$ as a parameter. We learn each light intensity $\mathbf{p_i}$ as a binary value to limit each light intensity to a positive and finite value, and to make light multiplexing robust. Gumbel softmax trick [8] is employed for this purpose while keeping the network differentiable.

The decoding layers is composed of a fully connected layer, and series of up-sampling layers. In each up-sampling layer, up-sampling by bilinear interpolation, convolution, batch normalization and activation (ReLU) are repeated twice.

4.2. Training LT encoder using synthetic data

To avoid preparation of a huge amount of ground truth dataset of LT, we prepare a dataset of synthetic LT for the

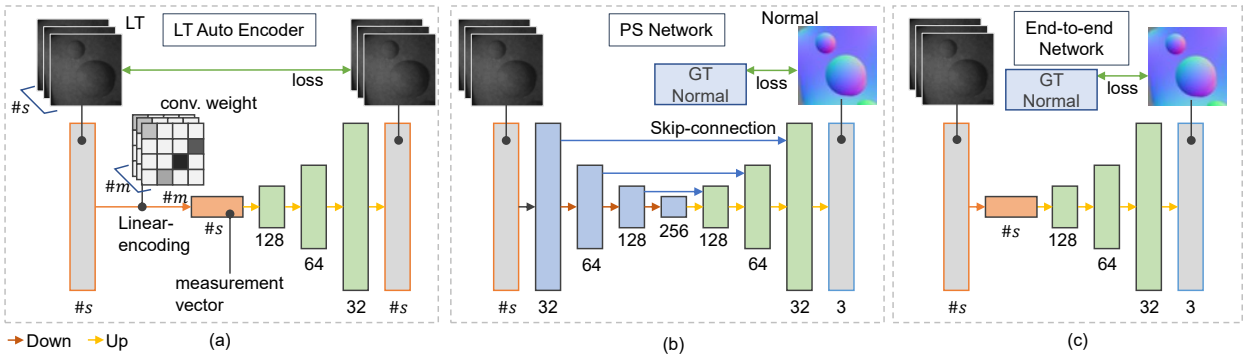


Figure 5. The structure of (a) LT auto-encoder, (b) PS network, and (c) end-to-end network. To realize PS from using multiplexed illumination, two independent network (a) + (b) or single network (c) for comparison in Sec. 5.4 are utilized.

training of the network. There are two rendering methods to synthesize LT, *i.e.*, rendering primal views from the camera, or dual views from the projector. Since rendering primal views requires computationally expensive rendering cost because of using optical diffuser, we synthesize dual views from the projector. In a synthetic scene, diffuser is approximated by a set of point light sources on a plane. The scene is rendered using a simple rasterizer which considers only direct diffuse reflections on the object's surface. For synthesis, we use 3d models from Blobby Shape Dataset [9] as 3D shapes and augment them into various positions, poses, and scales. L-1 norm between input and predicted LT is used for training.

Since the synthetic data generated by rasterization does not take into account the effects of global illumination nor ambient light in the real environment. To prevent artifacts during restoration, after training multiplexing pattern with the LT-auto-encoder on the synthetic data, we fine-tuned the weights of the LT-decoder using a few real measurements of real environment.

Since LT encoder is trained for specific configuration of a projector and a camera, if the configuration has changed, shape of LT also changes and retraining is required. Therefore, once setup is fixed, synthetic data are automatically generated based on calibration parameters for the setup, and the networks are retrained. Note that only manual task is taking a few images using the actual system for fine-tuning and it is not a heavy task.

4.3. Network architecture for robust PS

The PS network takes the shading images as input and estimates a surface normal. Since the shading images depends strongly on pixel coordinates under close light sources, UNet structure [18] is utilized to preserves the global position of a feature in the image. The network structure is shown in Fig. 5(b). It consists of three down-sampling layers and three up-sampling layers, which are connected by skip connections. In each layer, batch normalization (BN), convolution, and activation (ReLU) are performed twice in each layer. The loss function is defined as

$$L = \|n - \hat{n}\|_2^2, \tag{4}$$

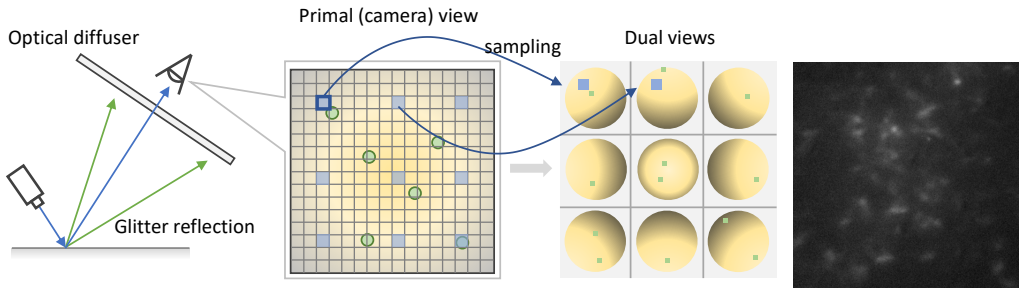
where n denotes the ground-truth (GT) normal and $\hat{n}$ denotes estimated normal.

Additionally, a structure of end-to-end network is shown in Fig. 5(c). This network is a slight modification of auto-encoder, where the number of channels in the output layer is changed for normal estimation. It takes sparse sample of LT as input and estimates a surface normal. In this network, the only loss is normal loss (Eq. (4)) for training, but shading images under arbitrary lighting conditions are not used.

4.4. Filtering angular distribution of reflected light

To handle complicated BRDF for PS, one solution is to apply filtering to light rays. When a single ray is reflected at a glitter surface, outgoing rays travel toward varying directions according to the surface's BRDF as shown in Fig. 6(a). These reflected rays form a unique pattern on the diffuser, where several strong intensity points are depicted as an angular distribution of reflected light. For example, we consider a BRDF which is represented by the combination of diffuse and sharp specular reflections as shown in Fig. 1(a). A corresponding observation should be a combination of a low-frequency component from the diffuse reflection and bright spots by glitter rays as shown in Fig. 6(b).

To estimate the surface normal, a component of diffuse reflection is extracted by applying a low-pass filter on the captured images, then, pixels are sampled for dual views. Note that since we assume dual photography for our PS, such simple process is effective for filtering light rays, which is not obvious for usual PS setups. We refer to this process as ray filtering. It can be implemented as Gaussian filtering on the imaging, or can be implemented in an optics, such like capturing in the out-of-focus setting.



(a) (b)

Figure 6. (a) Schematic illustration of light ray filtering for primal and dual view. (b) Example of an actually captured image. It is confirmed that glitter surface makes complicated reflection on captured image.

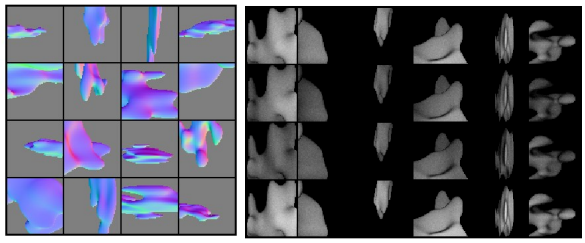


Figure 7. Example of rendered surface normals and shading images using Blobby Shape Dataset [9].

5. Experiments

5.1. Evaluation with synthesized data

In the synthetic scene, the diffuser and camera are placed at the same location at a distance of 5 from the center of the target object, and the projection is made on a 5x5 square area on the diffuser. The dataset consists of 8 blobs, each of which is randomly augmented with the object center position in a sphere of radius 1, free orientation, and size ranging from 50% to 150%, and the number of data is multiplied by 500. 90 % of the dataset is used for training and 10% for validation. Examples of synthesized data is shown in Fig. 7. The resolution of the camera is 32x32 and the resolution of the projector is 32x32. In the brute-force method, 1024 measurements are required to sample the entire LT.

Fig. 8(a) shows a dual view and normals of one of the LTs reconstructed using 16 multiplexed patterns. From the results, we can confirm that the LT auto-encoder successfully reconstructed the contours and the overall unevenness of the real view, although the high-frequency components were lost. However, when the normals are estimated from these dual views by PSN, low frequency normals are obtained. The results of qualitative evaluation are shown in table 1. This suggests that the multiplexing pattern is one of the reasons for the low frequency normals. The same can be said for the end-to-end network explained in Fig. 5(c).

LT Estimation		Normal estimation	
PSNR	SSIM	LT A.E.+PSN	End-to-end
22.68	0.674	0.062	0.784

Table 1. Quantitative evaluation.

Fig. 8(b) shows some of the multiplexed patterns learned by each network. It can be seen that the LT auto-encoder and the end-to-end network learn independent sets of binary patterns to extract the features of the scene. We also estimated the patterns with real-values without using the gambel softmax trick for comparison. In the generated patterns, we can observe lower contrast than the binary patterns and they are prone to be more sensitive to noise.

5.2. Setup of real data experiments

We validate the proposed method using real-world experiments. The prototype is shown in Fig. 9, where it consists of a LCD projector, a camera (Basler acA1920-155uc) and an optical diffuser (plastic), which is placed in front of the camera. The resolution of the projector is 1024×768 px and 1920×1080 px for the camera.

5.3. Glitter surface measurement with dual photography

First, we validate the effect of light ray filtering in the angular domain. A cube with a diffuse surface, the left half of which is covered with glitter paper (Fig. 9(b)) is measured. The LT is measured by brute-force manner, *i.e.*, turning on one pixel at a time out of 32×32 pattern. In the actual measurement, Hadamard code is used to compensate low SNR cause by low light intensity of the projector's single pixel illumination.

Figure 10 shows the effect of applying the filter explained in Sec. 4.4 to synthesize dual view that is relit by a virtual point light source from the center of the diffuser. The image is cropped so that the right half is the diffuse surface and the left half is the glitter material. Glitter is widely distributed in the left half of the image when no filter is applied. By applying a filter of larger size k_f, it weakens the glitter effect and makes the appearance closer to the diffuse

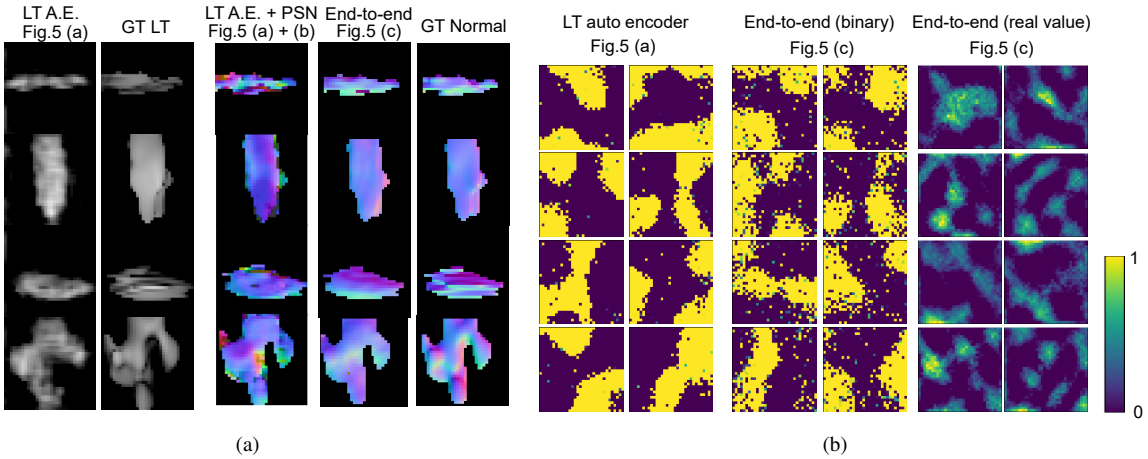


Figure 8. Verification of deep networks. (a) Result of dual view synthesis using LT encoder network and surface normal estimation using PSN. (b) Optimized patterns for multiplex lighting using the proposed networks.

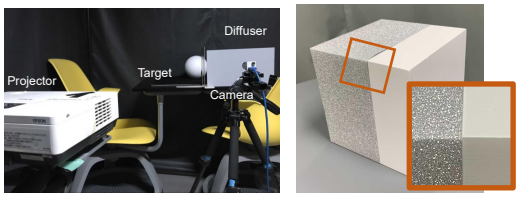


Figure 9. (a) Prototype for real-world measurement. (b) Target object. A cube with a diffuse surface, the left half of which is covered with glitter paper.

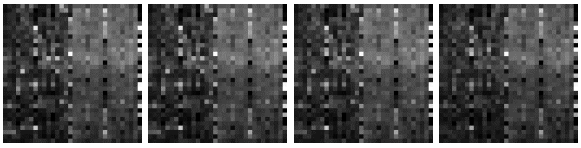


Figure 10. Result of ray filtering (contrast enhanced). Each Dual view is smoothed by a filter of a different size. Note that as the filter size increases, the glitter in the left half becomes weaker.

surface. It can also be confirmed that the diffuse surface is not affected by filtering.

Next, PS is applied to the dual view images with the filtering. To generate a shading image for PS, we sample four pixels from the captured image to synthesize dual views. Dual views are smoothed by ray filtering to reduce the effect of glitter pixels, then, the surface normals are estimated by PS using the calibrated light source positions. For comparison, surface normals are also estimated by using primal setup, *i.e.*, the optical diffuser is placed in front of the projector and illuminated by the single pixel projection, which makes a virtual point light source. After the captured images are smoothed by filtering (Gaussian kernel is used in the experiments), the surface normal is estimated. The object orientation is adjusted so that the primal and dual setups have the same viewpoint.

The results are shown in Fig. 11. The filter sizes are adjusted so that the standard deviation of the normals are the same for ray filter and spatial filter after smoothing (spatial: $8.13°$, proposed: $8.08°$). In the case of the primal setup, the edge of the two faces is blurred in surface normal image after the filtering, while it keeps sharp edge in the dual setup. In addition, high frequency errors are effectively decreased in dual setup, whereas low frequency artifacts appear in primal setup. Fig. 12 shows horizontal profiles of the surface normal (only y direction) of Fig. 11 with/without filtering in blue and red colors. From the results, it is confirmed that the proposed method preserves high-frequency shapes, whereas sharp edge is blurred out with primal setup.

5.4. Multiplexed measurement in real-world environment

We performed measurements using multiplexed patterns in a real environment. For actual projection, the pattern value is normalized between 0 to 255 accordingly. 16 patterns are utilized for the measurement. Shading images reconstructed from the multiplexed measurement is shown in Fig. 13. Although some areas have artifacts, the dual views are basically restored correctly, which proves that LT is reconstructed in the network. In the results, (a) and (b) are reconstructed by using a model trained with only synthesized data, whereas (c) and (d) are fine-tuned by real data. Since synthetic data cannot deal with noise and complicated illumination in the real environment, quality of generated shading images are much worse than those of (c) and (d).

The normal estimated by PSN is shown in Fig. 14. The reference surface normal is estimated from the LT reconstructed by brute-force measurements and they are shown in

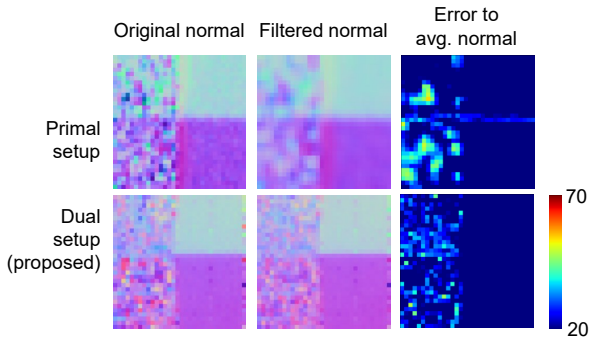
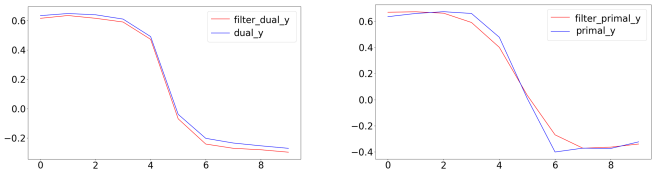


Figure 11. Surface normals estimated for glitter (left-half of the each image) and diffuse (right-half of the each image) surfaces. A spatial filter is applied to the primal setup (top), and an angular filter is applied to the dual setup (bottom). The normals are normalized so that each filtered normal has the same standard deviation. It is confirmed that dual setup decreases high frequency noises, but keeps sharp boundary edge, whereas primal setup blurred entire image



(a) Dual setup (b) Primal setup

Figure 12. Profiles of surface normal with and without filtering. It is shown that dual setup does not change the global shape, whereas the shapes of profiles are largely modified for primal setup.

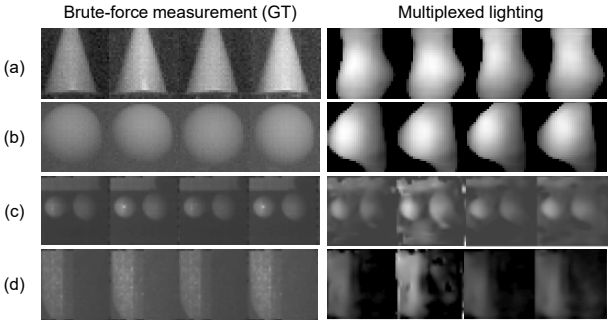


Figure 13. LT measurement and reconstruction with multiplexed illumination of four objects. (a) a cone, (b) a single ball, (c) two balls, and (d) a box with two planes. Each tiled image is a dual view synthesis result corresponding to four virtual illumination conditions. (a) and (b) are reconstructed by using models trained only with synthetic data, and (c) and (d) are reconstructed by model with fine-tuning using real data. It is clearly shown that the fine-tune significantly improves the results.

the second column. There are two balls in the ball scene (top row) and a two faces box in the board scene (bottom row). Since PS is trained using a synthesized dataset whose light-

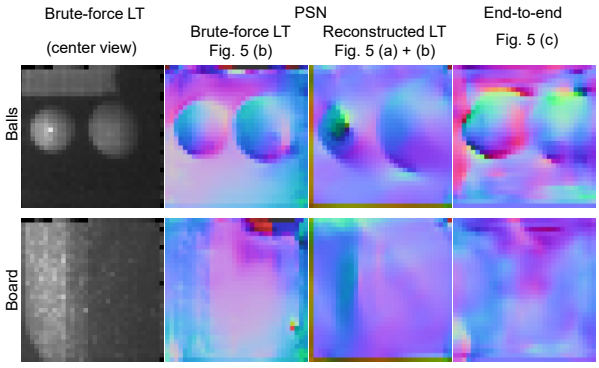


Figure 14. Surface normal estimation results by photometric stereo network with multiplexed measurement. It is shown that our joint network of LT auto-encoder and PSN (the 3rd column) is better than the end-to-end network (the 4th column). Please see the main text for more detail.

ing condition is as same as the real environment, estimated surface normals are expected to be same as the reference. When LT is estimated by multiplexed measurement (third column), the shape of the object is consistently estimated in most places, except left ball in the ball scene. Since the left ball has strong specularity, we think that the error is caused by such an unexpected BRDF.

The results of the end-to-end network (shown in Fig. 5(c)) are shown in 4th row for comparison. As shown in the figure, the spheres and the boards are reconstructed globally correct, however, some areas, such as the top of the board and right ball are mis-estimated. From the results, although some artifacts are observed, it is confirmed that the multiplexed measurement effectively worked.

6. Conclusion

We have tackled the PS measurement using the setup with a diffuser. We have proposed joint learning of the multiplexed lightings and reconstruction algorithm using DNN consists of a simple differentiable renderer and PS reconstruction modules. We have shown the multiplexed lighting can effectively reduce the number of measurement of LT acquisition. Also we have learned lighting pattern can deal with materials with complicated BRDF, such as glitter surfaces. We also proposed a photometric stereo method using diffuser and dual photography. Using our dual setup, we can estimate the normal of the glitter surface, where unpredictable reflections occur, by filtering the light rays in the angular domain. This effectiveness of the proposed method was verified in both synthetic and real-world experiments. In the future, conditional VAE is planned to generalize the configuration of the setup.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 20K19825, 20H00611, 18K19824, 18H04119 in Japan.

References

[1] Principles and prospects for single-pixel imaging, 2019. 2

[2] Svetlana Barsky and Maria Petrou. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. *PAMI*, 25(10), 2003. 2

[3] Guanying Chen, Kai Han, and Kwan Yee K. Wong. PS-FCN: A flexible learning framework for photometric stereo. In *ECCV*, 2018. 2

[4] Lixiong Chen, Yinqiang Zheng, Boxin Shi, Art Subpa-asa, and Imari Sato. A Microfacet-based Model for Photometric Stereo with General Isotropic Reflectance. *PAMI*, 2019. 2

[5] Wenzheng Chen, Parsa Mirdehghan, Sanja Fidler, and Kiriakos N. Kutulakos. Auto-Tuning Structured Light by Optical Stochastic Gradient Descent. In *CVPR*, 2020. 2

[6] Marco F. Duarte, Mark A. Davenport, Dharmpal Takhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2), 2008. 2

[7] Yoshio Hayasaki and Ryo Sato. Spectral imaging with a single pixel camera. In Abdul A. S. Awwal, Khan M. Iftekharuddin, and Mireya García Vázquez, editors, *Optics and Photonics for Information Processing XII*, volume 10751, pages 29 – 34. International Society for Optics and Photonics, SPIE. 2

[8] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 4

[9] Micah K. Johnson and Edward H. Adelson. Shape estimation in natural illumination. In *CVPR*, pages 2553–2560, 2011. 5, 6

[10] Kaizhang Kang, Zimin Chen, Jiaping Wang, Kun Zhou, and Hongzhi Wu. Efficient reflectance capture using an autoencoder. *TOG*, 37(4), 2018. 2

[11] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2

[12] Daisuke Miyazaki, Kenji Hara, and Katsushi Ikeuchi. Median photometric stereo as applied to the segonko tumulus and museum objects. In *IJCV*, volume 86, 2010. 2

[13] Yasuhiro Mukaigawa, Yasunori Ishii, and Takeshi Shakunaga. Analysis of photometric factors based on photometric linearization. *Journal of the Optical Society of America A*, 24(10), 2007. 2

[14] Thanh Trung Ngo, Hajime Nagahara, Ko Nishino, Rin ichiro Taniguchi, and Yasushi Yagi. Reflectance and Shape Estimation with a Light Field Camera Under Natural Illumination. *IJCV*, 127(11-12), 2019. 3

[15] Matthew O'Toole, Supreeth Achar, Srinivasa G. Narasimhan, and Kiriakos N. Kutulakos. Homogeneous codes for energy-efficient illumination and imaging. *ACM Trans. Graph.*, 34(4), 2015. 2

[16] Matthew O'Toole, Ramesh Raskar, and Kiriakos N. Kutulakos. Primal-dual coding to probe light transport. *ACM Trans. Graph.*, 31(4), 2012. 2

[17] Pieter Peers, Dhruv K. Mahajan, Bruce Lamond, Abhijeet Ghosh, Wojciech Matusik, Ravi Ramamoorthi, and Paul Debevec. Compressive light transport sensing. *ACM Trans. Graph.*, 28(1), 2009. 2

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *ECCV*, 2015. 5

[19] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep Photometric Stereo Network. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, volume 2018-Janua, 2017. 2

[20] Yoav Y. Schechner, Shree K. Nayar, and Peter N. Belhumeur. Multiplexing for optimal lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1339–1354, 2007. 1, 2

[21] Pradeep Sen, Billy Chen, Gaurav Garg, Stephen R. Marschner, Mark Horowitz, Marc Levoy, and Hendrik P.A. Lensch. Dual photography. *ACM Trans. Graph.*, 24(3):745–755, 2005. 2, 3

[22] Pradeep Sen and Soheil Darabi. Compressive dual photography. *Comput. Graph. Forum*, 28(2):609–618, 2009. 2

[23] Kobra Soltanlou and Hamid Latifi. Three-dimensional imaging through scattering media using a single pixel detector. *Applied Optics*, 58(28), 2019. 1, 2

[24] B. Sun, M. P. Edgar, R. Bowman, L. E. Vittert, S. Welsh, A. Bowman, and M. J. Padgett. 3D computational imaging with single-pixel detectors. *Science*, 340(6134), 2013. 1, 2

[25] Ming Jie Sun, Matthew P. Edgar, Graham M. Gibson, Baoqing Sun, Neal Radwell, Robert Lamb, and Miles J. Padgett. Single-pixel three-dimensional imaging with time-based depth resolution. *Nature Communications*, 7, 2016. 2

[26] Michael W. Tao, Jong Chyi Su, Ting Chun Wang, Jitendra Malik, and Ravi Ramamoorthi. Depth Estimation and Specular Removal for Glossy Surfaces Using Point and Line Consistency with Light-Field Cameras. *PAMI*, 38(6), 2016. 3

[27] Ting-chun Wang, U C Berkeley, Alexei A Efros, U C Berkeley, and Ravi Ramamoorthi. SVBRDF-Invariant Shape and Reflectance Estimation from Light-Field Cameras 1 Introduction 2 Related Work 3 Differential Stereo. *CVPR*, 2016. 3

[28] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ECCV*, 2011. 2

[29] Tai Pang Wu, Kam Lun Tang, Chi Keung Tang, and Tien Tsin Wong. Dense photometric stereo: A Markov random field approach. *PAMI*, 28(11), 2006. 2

[30] Yiwei Zhang, Matthew P. Edgar, Baoqing Sun, Neal Radwell, Graham M. Gibson, and Miles J. Padgett. 3D single-pixel video. *Journal of Optics (United Kingdom)*, 18(3), 2016. 2