



# A deep-shallow and global-local multi-feature fusion network for photometric stereo

Yanru Liu<sup>a</sup>, Yakun Ju<sup>a,\*</sup>, Muwei Jian<sup>b,c</sup>, Feng Gao<sup>a</sup>, Yuan Rao<sup>a</sup>, Yeqi Hu<sup>a</sup>, Junyu Dong<sup>a,\*</sup>

<sup>a</sup> Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China

<sup>b</sup> School of Information Science and Engineering, Linyi University, Linyi 276000, China

<sup>c</sup> School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China

## ARTICLE INFO

### Article history:

Received 18 October 2021

Received in revised form 13 December 2021

Accepted 19 December 2021

Available online 24 December 2021

### Keywords:

Photometric stereo

3D reconstruction

Deep neural networks

Convolutional neural network

Multi-feature fusion

## ABSTRACT

Recovering 3D surfaces based on the photometric stereo is a challenging task, due to the non-Lambertian surface of real-world objects. Although much effort has been made to address this issue, existing photometric stereo methods based on deep learning did not fully consider the influence of global-local features and deep-shallow features on the training process. How to combine multi-feature into a framework effectively to overcome their drawbacks has not been explored. Therefore, we propose a novel multi-feature fusion photometric stereo network (MF-PSN), focusing on both local-global and deep-shallow features fusion. Global-local feature fusion maintains the features under different illuminations and the most salient features of all illuminations, thereby effectively uses the information of each input image. Deep-shallow feature fusion keeps the features from deep and shallow layers with different receptive fields, which effectively improves the accuracy and robustness of the model. Experiments show that multi-feature fusion can make full use of the information of the input image to achieve a better reconstruction of surface normals of the object. Extensive ablation studies and experiments on the widely used DiLiGenT benchmark dataset have well verified the effectiveness of our proposed method. In addition, testing on the Gourd & Apple dataset and Light Stage Data Gallery verifies the generalization of our method.

© 2021 Elsevier B.V. All rights reserved.

## 1. Introduction

Recovering the 3D shape of an object is a pivotal problem in many computer vision applications [1,2]. Photometric stereo aims to recover the surface normals of an object through a set of input images, under different light directions [3,4]. Traditional non-Lambert photometric stereo methods mostly deal with real-world non-Lambert surfaces by approximating bidirectional reflectance distribution functions (BRDFs) [5–7] or rejecting the non-Lambertian outliers [8–10]. However, these methods have limited accuracy and can only be applied to limited materials.

Meanwhile, deep learning frameworks have shown potential abilities in many fields [11–16]. Recently, photometric stereo methods based on deep learning are proposed, which have good generalization ability [17–20]. Unfortunately, previous methods ignore multi-feature fusions. There are two main problems in the current methods. On the one hand, the previous full-pixel photometric stereo methods, such as PS-FCN [21], only select the most salient representation of each feature

from input images features. However, the global selection mechanism ignores some non-maximum features of images, while the discarded local features may still be important to the estimation of surface normals. On the other hand, the previous photometric stereo networks rarely employ the fusion of features from different network layers, i.e., the deep feature and shallow feature. Both deep and shallow features have an irreplaceable impact on network learning, due to the deep and shallow features with different receptive fields, may contain unique information, which are good for reconstructing the surface normals of the object.

In this work, we pay special attention to the combination of multi-features in a framework, focusing on solving the following challenges: (1) Combined global information and local information. The global and local features of images are crucial to the performance of 3D reconstruction. Nevertheless, in the existing algorithms, the combination of global features and local features is seldom considered. Therefore, how to combine global and local information into a network framework to reinforce its performance is a tricky task. (2) Deep-shallow features fusion. Deep and shallow features retain different information, so it is important to explore the impact of deep and shallow features fusion on network learning, so as to design a more reasonable network structure. (3) The robustness of a different number of input images. Previous

\* Corresponding authors.

E-mail addresses: [juyakun@stu.ouc.edu.cn](mailto:juyakun@stu.ouc.edu.cn) (Y. Ju), [dongjunyu@ouc.edu.cn](mailto:dongjunyu@ouc.edu.cn) (J. Dong).

methods usually fail to maintain both the best performance when meets dense input images and sparse input images, which leads to the instability of the results.

To address the aforementioned issues, we propose a novel solution to investigate the photometric stereo, namely MF-PSN, which integrates global-local features and deep-shallow features into a network framework. The multi-features fusion network structure is illustrated in Fig. 1. Firstly, max-pooling is used for feature fusion, which extracts the global features. In order to preserve more information in the original image, our network fuses local features and global features in the shallow layer. Specifically, we first use max-pooling operation for shallow global feature fusion. At the same time, we use catenate operation to fuse the local features obtained by the first feature extractor with the global features obtained by the max-pooling operation. We believe that the shallow layer of the network can retain more original information. Then, the features passing through the second feature extractor are fused in a deeper layer, and only the global features are fused at this time. Two different levels of feature fusion can operate on different receptive fields to obtain richer information. Furthermore, using deep and shallow multi-layer feature fusion (twice max-pooling operations in different layers) can jointly optimize network learning and improve network robustness. Extensive ablation studies and experiments are conducted on the widely used DLiGenT benchmark dataset [22], which demonstrates the effectiveness of our method. Additionally, testing on the Gourd & Apple dataset and Light Stage Data Gallery dataset verify the generalization of our method.

## 2. Related work

The traditional photometric stereo method was proposed by Woodham in 1980 [3] based on the assumption of Lambertian. However, there are not always ideal Lambertian surfaces in the real world, therefore, many methods were proposed to deal with non-Lambertian surface problems. Based on the strategy of rejecting outliers, the non-Lambertian areas are regarded as the outliers and being removed [8–10,23]. However, these methods can only deal with surfaces which exist local or sparse non-Lambertian regions. The analytic reflectance model-based strategy is to perform complex modeling on the reflection properties of the surface of the object to form an equation for solving the surface normal and material coefficients of the object, thereby solving the reflection coefficient of the object surface [24–27]. This method can realistically restore the surface material of the object by relight rendering getting the re-illumination image as close to the real image as possible if get the accurate illumination model coefficients when solving the normal direction of the surface. The strategy based on the general

properties of BRDFs are dedicated to finding a simple and accurate bi-directional reflectance distribution function that can express the object surface [5–7,28–30]. However, these methods can only handle limited non-Lambertian surfaces and consume a lot of calculations.

Recently, deep learning has been widely introduced into many research areas [31–37]. On the field of surface normal reconstruction, DPSN introduced the deep learning technology into the photometric stereo field for the first time [17]. By using the dropout function, DPSN randomly discards some values to replace the shadow and specular reflection in the real world, which is similar to the outlier rejection method of traditional methods. It uses end-to-end learning to learn the surface normal of objects directly from images. Compared with traditional methods, its performance is improved. However, its application is limited by the predefined light directions. In order to solve this problem, two kinds of methods are proposed: One is the pixel by pixel method and the other is the full pixel method.

The observation map is used to solve the problem of order-agnostic, namely the pixel by pixel method, which was proposed in CNN-PS [19]. Later, LMPS [38] also uses an observation map to deal with the problem of arbitrary input. Different from CNN-PS, LMPS uses an occlusion layer to simulate the effect of shadow, which can improve the performance of the method in shadow areas. However, these methods do not focus on the fusion of global-local features and deep-shallow features.

Input any number of images into the network, and then the arbitrary input images are converted into a fixed quantity feature is called the full pixel method. PS-FCN [21] uses a max-pooling operation to extract the global features of the network and then inputs them into the regression network. Later, Manifold-PSN [39] was proposed, which uses manifold mapping to handle the high-dimensional multi-features. Furthermore, Ju et al. proposed Attention-PSN to better handle the complex structure [40]. However, these methods do not consider the preservation of global-local features and the fusion of deep-shallow features, which makes the limited accuracy.

In addition, some other methods have been proposed for the deep photometric stereo. For example, Tatsunori Taniai and Takanori Maehara apply unsupervised strategy to photometric stereo [41], which optimizes the model by the reconstruction loss between re-rendered images and input images. Ju et al. further apply both reconstruction loss and normal loss to optimize the photometric stereo network, namely DR-PSN, to form a closed-loop structure and improve the estimation of surface normals [42]. GPS-Net [43] combines the full pixel method with the pixel by pixel method to construct a photometric stereo network based on graph convolution [44,45]. The authors input the images into the pixel by pixel network based on graph convolution network to effectively solve the problem of sparse illumination, which is

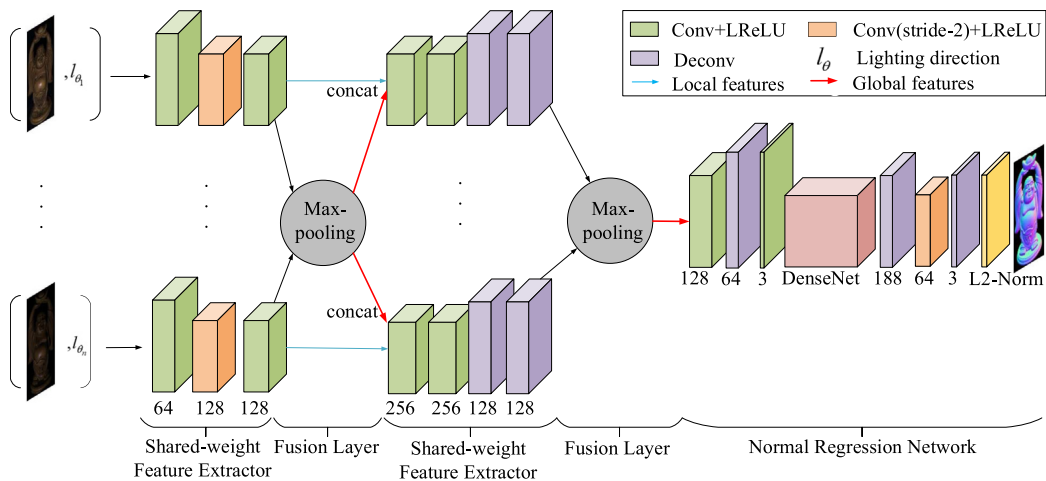


Fig. 1. Network structure (the number under each layer means the channel of the convolution, where the kernel is  $3 \times 3$ ).

different from CNN-PS [19]. Then input the extracted features into the full pixel network, and the normal map of the object is obtained by regression. In this work, the authors consider the spatial information of features but not global-local and deep-shallow multi-fusion.

In summary, these advanced methods do not pay attention to the fusion of global-local and deep-shallow features at the same time, so they can not deal with complex structural areas well. On the contrary, we propose a multi-fusion framework, which adopts the full pixel method, fully considers the fusion of global-local and deep-shallow features, and can effectively use a variety of information of the input image to better deal with the complex structure area.

### 3. Proposed method

We propose a network model of deep-shallow and global-local multi-fusion, as shown in Fig. 1. The max-pooling operation has been proved to solve the problem of order-agnostic by PS-FCN [21], which fuses features on the arbitrary number of input extraction features with a random order automatically reserves the most significant features. Therefore, our network structure adopts this strategy and uses two max-pooling for deep and shallow multi-layer feature fusion. The first max-pooling operation is used to extract features from both global and local. In this way, the original local feature information is retained with the global feature, which is conducive to more sufficient feature learning and training and avoids missing useful features of images non-maximum features. We argue that in the early stage of training, we should keep as much original information as possible, while discarding local features too early will lose useful information and reduce the accuracy of the results. We showed the network performance after adding local information fusion is better than the network model that only retains the global features (as tabulated in Table 2). The second max-pooling only extracts global features. This is because sufficient feature learning has been done before, and just extracting global features can improve the efficiency of network operation. Furthermore, the deep-shallow feature fusion can fuse the features with different receptive fields, which improves the accuracy of results. Concretely, the network architecture consists of six convolution layers, two downsampling layers, five upsampling layers, two max-pooling layers, one 24-layer DenseNet module (the structure is shown in Table 1), and one L2-Norm layer, which is used for normalizing the estimated surface normal vectors.

Our network is trained with NVIDIA GeForce GTX 1080 with Adam optimization. The learning decay is set to 0.5, learning rate = 0.001, regularization parameters  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ . The network uses LeakyRelu as the activation function. Our network uses the torch.nn.init.kaiming normal initializes the weight and shared-weight training on images. We train the model using a batch size of 32 for 30 epochs.

**Table 1**  
The structure of the DenseNet layer used in this paper.

Layers	Output Size	DenseNet-24(k = 32)
Convolution	64×64	7×7 conv, stride 1
Pooling	32×32	3×3 maxpool, stride 2
Dense Block(1)	32×32	1×1 conv
	32×32	3×3 conv
Transition Layer(1)	32×32	1×1 conv
	32×32	1×1 average pool, stride 1
Dense Block(2)	32×32	[1×1 conv]×2
	32×32	[3×3 conv]×2
Transition Layer(2)	32×32	1×1 conv
	32×32	1×1 average pool, stride 1
Dense Block(3)	32×32	[1×1 conv]×4
	32×32	[3×3 conv]×4
Transition Layer(3)	32×32	1×1 conv
	32×32	1×1 average pool, stride 1
Dense Block(4)	32×32	[1×1 conv]×3
	32×32	[3×3 conv]×3

The default number of input images for training is 16. We choose cosine similarity loss as the loss of the network, as follows:

$$\text{loss} = \frac{1}{HW} \sum_i (1 - N_i \cdot \tilde{N}_i), \quad (1)$$

where  $HW$  is the resolution,  $i$  is the index of a pixel of the image,  $N_i$  is the ground-truth, and  $\tilde{N}_i$  is the estimated surface normal.

## 4. Experimental evaluation

### 4.1. Datasets

In our experiment, the Blobby dataset [46] and the Sculpture dataset [47], rendered with MERL BRDFs dataset [48], are used for training. The evaluation of the model is based on the DiLiGenT dataset [22]. This dataset contains ten objects with 96 different light directions. In our experiment, the average angle error is used as the evaluation index of the model. The smaller the value is, the better the model is. The calculation formula is as follows:

$$\text{MAE} = \frac{1}{HW} \sum_i \arccos(N_i \cdot \tilde{N}_i). \quad (2)$$

In addition, we also use two other real-world datasets, the Gourd&Apple dataset [49] and Light Stage Data Gallery dataset [50]. The Gourd&Apple dataset contains three different objects, namely Apple, Gourd1, and Gourd2. There are 112 images of object Apple, 102 images of object Gourd1, and 98 images of object Gourd2, with different light directions. The Light Stage Data Gallery dataset contains six objects with strong non-Lambertian surfaces. We used three objects, namely helmet, knight fighting, and plant. There are 253 images with different light directions for each object. The two datasets provide images without the ground-truth of surface normal. Therefore, the experimental results on these two datasets can only be evaluated qualitatively.




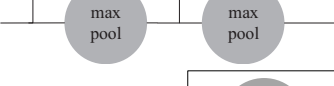
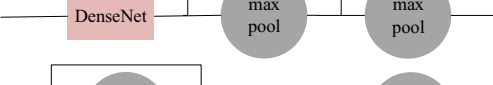

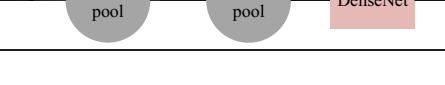
### 4.2. Network analysis

We quantitatively analyze our network structure and test different network structures on the DiLiGenT dataset [22], with 96 input images. The experimental results of average MAE of DiLiGenT dataset are shown in Table 2, and Fig. 2 shows some visualized examples in the DiLiGenT dataset. First, we verify the effectiveness of global-local fusion, and then we discuss the effectiveness of deep-shallow fusion. ID (0) is with the same architecture of PS-FCN [21]. ID (1) adds a max-pooling layer to fuse global features, in other words, two max-pooling are used for global feature extraction. ID (2) fuses the local features and global features extracted by ID (0) and then inputs them into the normal regression network. ID (3) adds global features extraction on the basis of ID (2), i.e., after fusing local features and global features, it extracts global features and inputs them into the normal regression network to fuse deep and shallow features. Furthermore, IDs (4), (5), and (6) add the 24-layer DenseNet module (as tabulated in Table 1) in different positions, to generate the estimated surface-normal.

#### 4.2.1. Effectiveness of global-local features fusion

It can be seen that the MAE of ID (0) is 8.39, while the MAE of ID (2) is 8.29, in Table 2, which is 0.1 lower than that of ID (0). The ID (2) uses global and local feature fusion, while the ID (0) only considers global feature fusion. Global fusion using max-pooling, which only retains the maximum of all features, leads to lacking some features as same significance as the maximum. Experiments show that the network using global and local features fusion can reduce the loss of important features and achieve better results. As shown in Fig. 2, the global-local features fusion clearly improves the accuracy of shadow regions, such as the head of the object "Cat" and the cloth of the object "Reading".

**Table 2**  
The effectiveness of different network structures.

ID	Method	MAE (DiLiGenT [22])
(0)		8.39
(1)		9.15
(2)		8.29
(3)		8.12
(4)		8.21
(5)		7.87
(6)		7.27

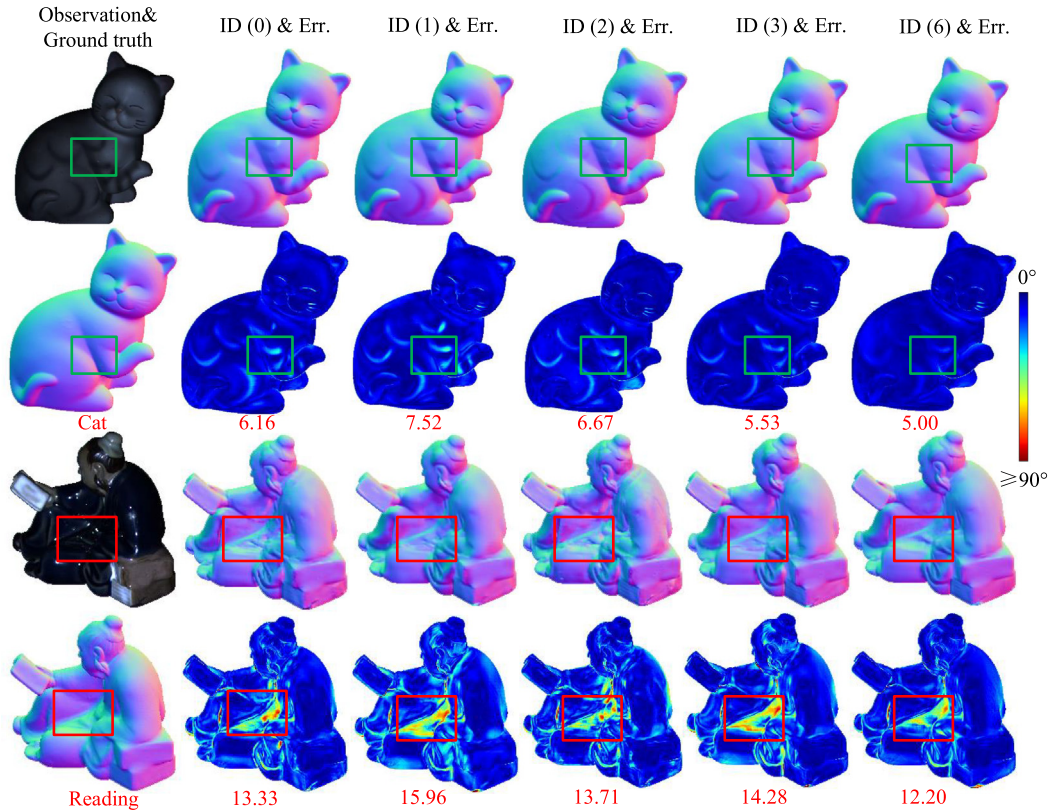
#### 4.2.2. Effects of deep-shallow fusion

It can be seen that the MAE of ID (1) is 9.15, while the MAE of ID (0) is 8.39. It shows that there may be no relationship between the number of max-pooling operations and the quality of the model. Two max-pooling operations even are worse than one counterpart. This is because two max-pooling for global feature extraction causes a large amount of local information loss, which makes it difficult to achieve a better result. However, the MAE of ID (3) is 8.12, which is 0.17 lower than that of ID (2). Compared with ID (2), ID (3) has one more deep features fusion. It shows that the deep and shallow multi-layer features fusion effect is better under the condition of the reasonable design, due to its various receptive fields and sizes. As shown in Fig. 2, the estimations of the cast shadows regions and the spatially-varying materials regions are much better in IDs (3) and (6).

Furthermore, as tabulated in Table 2, it can be seen that the position of the DenseNet module is important to the accuracy of the estimated surface normal. We argue that a very deep feature is not needed for the extractor of photometric stereo, because the per-pixel estimated surface normal is more like a low-level generation. Compared with IDs (4) and (5), ID (6) (our default setting) treats the DenseNet as a part of the regressor to decode the aggregated feature, which may benefit the prediction.

#### 4.2.3. Effects of different size of input images

In this section, we discuss the impact of different sizes of input images in training. Due to the memory of GPU, we can only train the size smaller than  $32 \times 32$ . We show the comparisons with different sizes in Table 3. We trained our MF-PSN with input images of sizes  $8 \times 8$ ,  $16 \times 16$ , and  $32 \times 32$ , respectively, and tested the results on the DiLiGenT benchmark dataset [22], with 96 input images.



**Fig. 2.** Visualized examples of DiLiGenT benchmark dataset. The first row of each sample represents the estimated normal maps, while the second row represents the error maps, based on the different network settings. The values represent MAE in degrees. Compared with IDs (0), (1), (2), and (3), the default settings ID (6) achieves better performance for surfaces with spatially-varying materials (green boxes) and cast shadows (red boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Table 3**

Comparisons of the results with different sizes of input images.

Method	MAE (DiLiGenT [22])
MF-PSN: $8 \times 8$	9.88
MF-PSN: $16 \times 16$	7.95
MF-PSN: $32 \times 32$	<b>7.27</b>

The experimental results are shown in Table 3. Intuitively, the larger the size of the image, the more texture and contextual information it contains, and the better features it can capture. In addition, when the image becomes larger, some discriminative features are better obtained. The larger the training image size, the better the result obtained. This is because larger images can provide more useful information for network learning, thereby improving the accuracy of network prediction results. However, if the size of the training input image is too large, the training time will increase dramatically and need GPUs with large memory. We argue that an appropriate larger image size (e.g.  $32 \times 32$ ) will improve the accuracy of the result.

#### 4.3. DiLiGenT benchmark comparisons

We compare our proposed MF-PSN with calibrated photometric stereo methods on the DiLiGenT dataset [22]. The methods we compared include traditional photometric stereo methods (represented by the first letter of the authors' name + published year) and photometric stereo methods based on deep learning, tested with 96 images, such as DPSN [17], PS-FCN [21], PS-FCN(Norm.) [51], CNN-PS [19], Attention-PSN [40], DR-PSN [42], IRPS [41], GPS-Net [43], CHR-PSN [52], LMPS [38], and SPLINE-Net [53]. The test results are shown in Table 4, and Fig. 3 shows visual comparisons (Goblet, Pot2, Cow, Buddha) of our method and other state-of-the-art methods, as well as the baseline (least square [3]).

In fact, many practical applications involve sparse photometric stereo. We, therefore, evaluate our MF-PSN with different numbers of input images. As tabulated in Table 5, we show the MAE of our MF-PSN, and recent state-of-the-art deep learning-based methods which can handle sparse input images, tested with only 10 input images. It is noting that the methods LMPS [38] and SPLINE-Net [53] are specially designed for sparse condition, which are trained with 10 input images, while other deep learning-based methods are trained with more input images (e.g., 16 or 32).

**Table 4**

Comparison of the calibrated photometric stereo methods on the DiLiGenT benchmark dataset, tested with 96 dense input images.

Method	Ball	Cat	Pot1	Bear	Pot2	Buddha	Goblet	Reading	Cow	Harvest	Avg.
Baseline [3]	4.10	8.41	8.89	8.39	14.65	14.92	18.50	19.80	25.60	30.62	15.39
WG10 [10]	2.06	6.73	7.18	6.50	13.12	10.91	15.70	15.39	25.89	30.01	13.35
AZ08 [49]	2.71	6.53	7.23	5.96	11.03	12.54	13.93	14.17	21.48	30.50	12.61
GC10 [54]	3.21	8.22	8.53	6.62	7.90	14.85	14.22	19.07	9.55	27.84	12.00
IA14 [6]	3.34	6.74	6.64	7.11	8.77	10.47	9.71	14.19	13.05	25.95	10.60
ST14 [5]	1.74	6.12	6.51	6.12	8.78	10.60	10.09	13.63	13.93	25.44	10.30
SPLINE-Net [53]	4.51	6.49	8.29	5.28	10.89	10.36	9.62	15.50	7.44	17.93	9.63
EW20 [55]	1.58	6.30	6.67	6.38	7.26	13.69	11.42	15.49	7.80	18.74	9.53
DPSN [17]	2.02	6.54	7.05	6.31	7.86	12.68	11.28	15.51	8.01	16.86	9.41
CK18 [56]	1.50	5.74	6.24	4.97	8.64	8.86	10.00	11.44	11.33	21.90	9.06
IRPS [41]	<b>1.47</b>	5.44	6.09	5.79	7.76	10.36	11.47	11.03	6.32	22.59	8.83
LMPS [38]	2.40	6.11	6.54	5.23	7.48	9.89	8.61	<b>13.68</b>	7.98	16.18	8.41
PS-FCN [21]	2.82	6.16	7.13	7.55	7.25	7.91	8.60	13.33	7.33	15.85	8.39
Attention-PSN [40]	2.93	6.14	6.92	<b>4.86</b>	6.97	7.75	8.42	12.90	6.86	15.44	7.92
DR-PSN [42]	2.27	5.42	7.08	5.46	7.21	7.84	8.49	12.74	7.01	15.40	7.90
GPS-Net [43]	2.92	5.42	6.04	5.07	7.01	7.77	9.00	13.58	6.14	15.14	7.81
CHR-PSN [52]	2.26	5.97	<u>7.04</u>	6.35	6.76	7.15	8.32	12.52	6.05	15.32	7.77
CNN-PS [19]	2.20	<b>4.60</b>	<b>5.40</b>	8.30	<b>6.00</b>	<u>7.90</u>	<b>7.30</b>	12.60	<u>8.00</u>	14.00	7.60
PS-FCN(Norm.) [51]	2.67	4.76	6.17	7.72	7.15	7.53	7.84	<b>10.92</b>	6.72	<b>12.39</b>	7.39
MF-PSN(Ours)	2.07	<u>5.00</u>	7.20	5.83	6.81	<b>6.88</b>	<u>7.46</u>	12.20	<b>5.90</b>	<u>13.38</u>	<b>7.27</b>

The value represents the MAE of the estimated surface normals, where the **bold** numbers indicate the best results, while the underlined values represent the second-best performance.

As tabulated in Table 4, the average MAE on DiLiGenT dataset [22] of our method (MF-PSN) is 7.27. Compared with the previous methods, our MF-PSN achieves the best or second-best performance on five objects. Our method especially performs well on objects with strong non-Lambertian surface or complex structure, such as "Goblet", "Cow", and "Buddha", as shown Fig. 3. It can be observed that our method can more accurately recover the surface normals in those regions with cast shadows (red boxes), such as the "shoulder" of the object "Buddha". Our method can achieve more accurate estimation in these regions, compared with other deep learning-based methods.

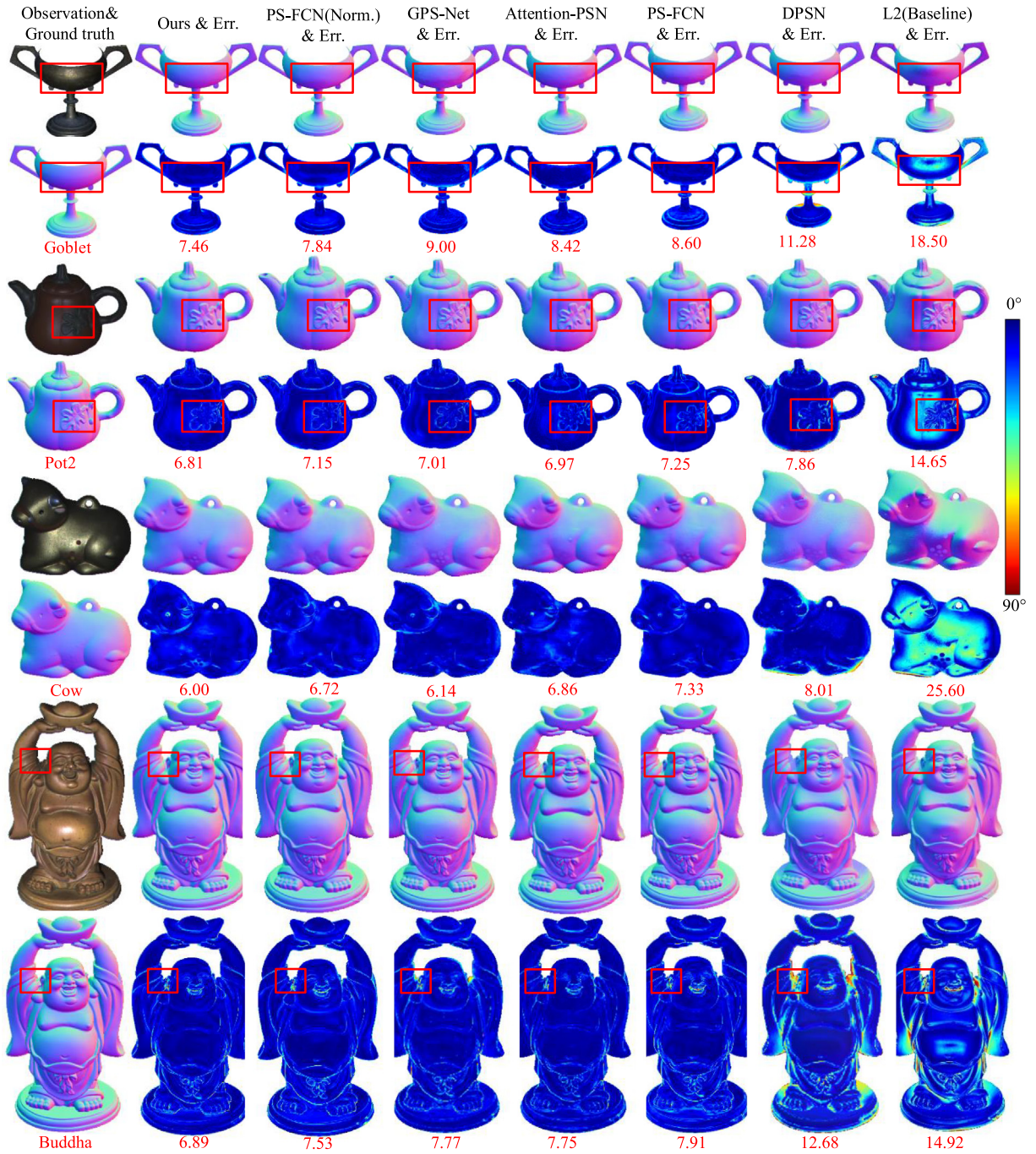
In addition, as tabulated in Table 5, our method achieves the best performance on the sparse input condition, and the prediction results of most objects are the best or second-best. Furthermore, we show the performance on the DiLiGenT benchmark with different numbers of the input images, as shown in Fig. 4, compared with the state-of-the-art methods [19,21,38,41,43,53]. It can be seen that our method achieves the smallest angular error, averaged over ten objects, with all the conditions for different numbers of input images. Moreover, our MF-PSN improves the performance significantly when meets the sparse inputs (e.g., 8, 16, 32). It illustrates the robustness of our MF-PSN, under sparse to dense input images, which illustrates the robustness and effectiveness of our multi-feature fusion structure.

#### 4.4. Evaluation with different numbers of input images

We compare the results of our MF-PSN and PS-FCN [21], with different numbers of input images in training and testing, to evaluate the robustness of our method. Specifically, we train our proposed method and PS-FCN with 8, 16, 32, 48 input images, and test with 16, 32, 64, and 96 input images of DiLiGenT dataset. The results are shown in Fig. 5.

As shown in Fig. 5, our method outperforms PS-FCN [21] on all the conditions. Furthermore, in the case of different numbers of test images, our method has fewer error fluctuations (standard error) than PS-FCN, which proves the effectiveness and robustness of our network.

Moreover, it can be seen that the best performance of MF-PSN exists in trained with 16 input images, while PS-FCN exists in 32 input images. The reason why our MF-PSN needs fewer input images than PS-FCN may be that the better deep-shallow feature extraction module and the global-local feature fusion mechanism can learn the powerful feature from less input images. Also, we observe that the performance of both MF-PSN and PS-FCN can not achieve the peak when input 48 images or 8 images. This could be explained by the three facts: (1) Each training sample has 64 images with different light directions in total.



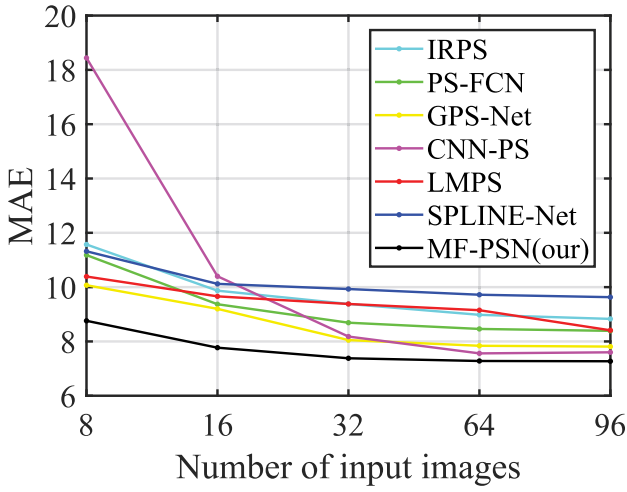
**Fig. 3.** Quantitative comparison on the DiLiGenT benchmark dataset, with 96 input images. Compared with PS-FCN (Norm.), GPS-Net, Attention-PSN, PS-FCN, DPSN, and L2 baseline method, our MF-PSN achieves better performance for surfaces with cast shadows (red boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 5**

Comparison of the calibrated photometric stereo methods on the DiLiGenT benchmark dataset, tested with 10 sparse input images.

Method	Ball	Cat	Pot1	Bear	Pot2	Buddha	Goblet	Reading	Cow	Harvest	Avg.
Baseline [3]	4.58	8.90	9.59	9.84	15.65	16.02	19.23	19.37	26.48	31.32	16.10
CNN-PS [19]	8.21	9.00	12.79	11.89	15.04	13.39	15.74	16.07	13.83	19.36	13.53
LMPS [38]	3.97	6.69	7.30	8.73	9.74	11.36	10.46	14.37	10.19	17.33	10.01
SPLINE-Net [53]	4.96	7.52	8.77	5.99	11.79	10.07	<u>10.43</u>	16.13	<u>8.80</u>	19.05	10.35
IRPS [41]	<b>2.12</b>	6.58	7.14	6.92	9.61	11.41	<u>14.99</u>	13.70	8.87	26.55	10.79
PS-FCN [21]	4.35	8.24	8.38	5.70	10.37	10.54	11.21	14.34	9.97	18.82	10.19
PS-FCN(Norm.) [51]	4.38	6.30	<b>7.05</b>	<u>5.92</u>	11.91	8.98	10.96	13.23	14.66	18.04	10.14
GPS-Net [43]	4.33	<u>6.81</u>	7.50	6.34	<b>8.38</b>	8.87	10.79	<u>15.00</u>	9.34	<u>16.92</u>	<u>9.43</u>
MF-PSN(Ours)	<u>2.97</u>	<b>5.55</b>	7.21	<b>4.89</b>	<u>9.16</u>	<u>7.43</u>	<b>9.87</b>	<b>12.92</b>	<b>8.41</b>	<u>16.39</u>	<b>8.48</b>

The value represents the MAE of the estimated surface normals, where the **bold** numbers indicate the best results, while the underlined values represent the second-best performance.



**Fig. 4.** Comparisons on different numbers of input images. Our method achieves the best performance under all the different numbers of input images.

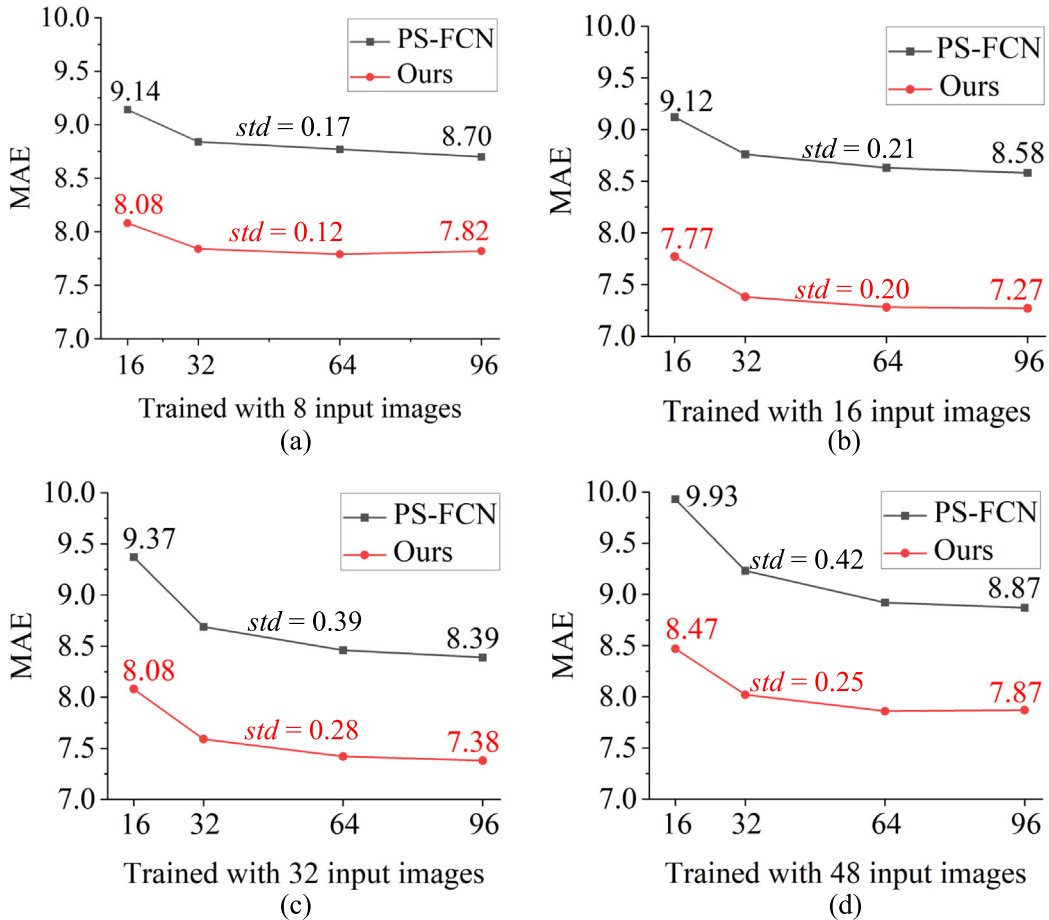
When trained with 48 input images randomly, the diversity and randomness of training data is worse than trained with 32 and 16 input images, which may impact the performance of the networks. (2) The max-pooling framework only keeps the maximum value from all the input features, therefore, too many input images may cause excessive loss of

useful information, which impacts the stability of networks convergence. (3) Too less input images (e.g., 8) will cause the loss of key features, which will mislead the pattern the trained networks.

#### 4.5. Evaluation on other datasets

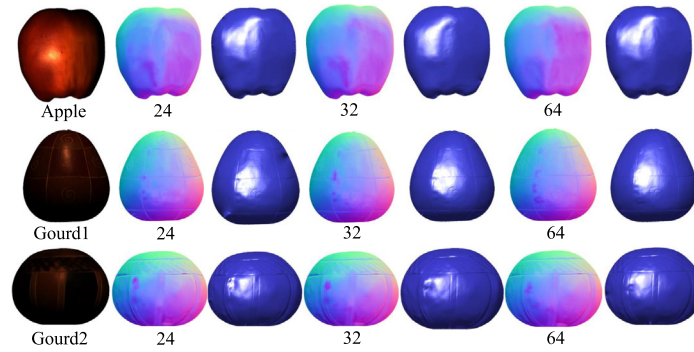
In order to verify the effectiveness and generalization of our method, we test our model on the Gourd & Apple dataset [49] and the Light Stage Data Gallery [50] dataset. “Apple”, “Gourd1”, and “Gourd2” are tested from the Gourd & Apple dataset. “Helmet”, “Knight fighting”, and “Plant” are tested on the Light Stage Data Gallery. Similarly, our method is trained with 32 input images, and show the results of the Gourd & Apple dataset tested with 24, 32, and 64 test images, in Fig. 6 and the results of the Light Stage Data Gallery dataset results tested with 24, 64, 96, 128, and 192 test images, in Fig. 7. Since there is no ground truth in these two datasets, we can only show the qualitative results.

As shown in Figs. 6 and 7, our proposed method recovers the surface normals of the object accurately. It can be seen our method achieves good performance when dealing with complex areas, such as the crinkles of Gourd1, and the screws of Helmet. Note that the material of the object Plant are never seen under training, while our method can achieve quite visually accurate results, which also illustrates the robustness of our MF-PSN. However, we also find that the estimated surface normal of the object “Knight\_fighting” is not very good, especially when the number of input images is sparse. We argue that the object “Knight\_fighting” is too large for the light to cover its entire area, which caused serious errors when using a small number of images.

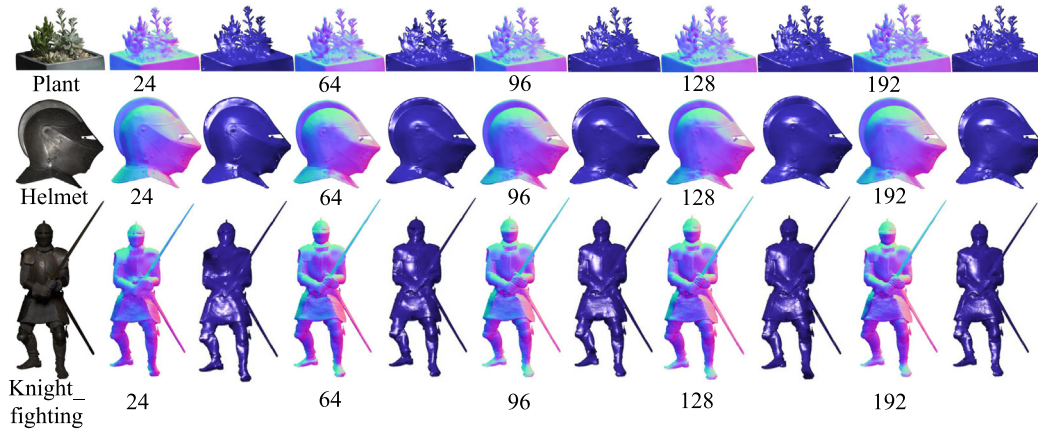


**Fig. 5.** Comparison of the results of different numbers of input images. (a), (b), (c), and (d) represent our MF-PSN and PS-FCN [21] trained with 8, 16, 32, and 48 input images, respectively. Std means the standard deviation.





**Fig. 6.** Qualitative results of our MF-PSN on the Gourds & Apple dataset. Due to the lack of ground truth, we further show the 3D reconstruction results of our estimated surface normal maps using [57], to clearly show the details.



**Fig. 7.** Qualitative results of our MF-PSN on the Light Stage Data Gallery dataset.

## 5. Conclusions

In this paper, we have proposed a deep-shallow and global-local multi-feature fusion network (MF-PSN) for calibrated photometric stereo. We discussed the effectiveness of fusion of deep-shallow and global-local features, with adequate ablation experiments. Our experiments show that our method achieves the state-of-the-art performance on the DiLiGenT benchmark test, with an MAE of 7.27. Meanwhile, the visual comparison has shown the ability of our method to deal with complex structural areas, which obtained promising results and clear details. And our method has smaller fluctuations and maintains the best performance when processing different numbers of training and verification images, which proves the robustness of our proposed method.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the National Key Scientific Instrument and Equipment Development Projects of China (Grant No.41927805), and the National Natural Science Foundation of China (Grants No.61501417 and No.61976123), the Royal Society - K. C. Wong International Fellowship (NIF\R1\180909) and the Taishan Young Scholars Program of Shandong Province.

## References

- [1] Z.-L. Sun, K.-M. Lam, Depth estimation of face images based on the constrained lca model, *IEEE Trans. Inform. Forensic Secur.* 6 (2011) 360–370.
- [2] W.-Z. Nie, A.-A. Liu, S. Zhao, Y. Gao, Deep correlated joint network for 2-d image-based 3-d model retrieval, *IEEE Trans. Cybern.* (2020).
- [3] R.J. Woodham, Photometric method for determining surface orientation from multiple images, *Opt. Eng.* 19 (1980) 139–144.
- [4] M. Jian, J. Dong, M. Gong, H. Yu, L. Nie, Y. Yin, K.-M. Lam, Learning the traditional art of chinese calligraphy via three-dimensional reconstruction and assessment, *IEEE Trans. Multimed.* 22 (2019) 970–979.
- [5] B. Shi, P. Tan, Y. Matsushita, K. Ikeuchi, Bi-polynomial modeling of low-frequency reflectances, *IEEE Trans. Pattern Anal. Mach. Intell.* 36 (2013) 1078–1091.
- [6] S. Ikehata, K. Aizawa, Photometric stereo using constrained bivariate regression for general isotropic surfaces, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE 2014, pp. 2179–2186.
- [7] S. Li, B. Shi, Photometric stereo for general isotropic reflectances by spherical linear interpolation, *Opt. Eng.* 54 (2015), 083104.
- [8] S. Barsky, M. Petrou, The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2003) 1239–1252.
- [9] F. Verbiest, L. Van Gool, Photometric stereo with coherent outlier handling and confidence estimation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE 2008, pp. 1–8.
- [10] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, Y. Ma, Robust photometric stereo via low-rank matrix completion and recovery, *Proceedings of the Asian Conference on Computer Vision*, Springer 2010, pp. 703–717.
- [11] J. Xiao, W. Jia, K.-M. Lam, Feature redundancy mining: Deep light-weight image super-resolution model, *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE 2021, pp. 1620–1624.
- [12] W.-Z. Nie, W.-W. Jia, W.-H. Li, A.-A. Liu, S.-C. Zhao, 3d pose estimation based on re-inforce learning for 2d image-based 3d model retrieval, *IEEE Trans. Multimed.* 23 (2020) 1021–1034.
- [13] C. Wang, Y. Wu, Z. Su, J. Chen, Joint self-attention and scale-aggregation for self-calibrated deraining network, *Proceedings of the 28th ACM International Conference on Multimedia 2020*, pp. 2517–2525.



- [14] C. Wang, X. Xing, Y. Wu, Z. Su, J. Chen, Dcsfn: Deep cross-scale fusion network for single image rain removal, *Proceedings of the 28th ACM International Conference on Multimedia 2020*, pp. 1643–1651.
- [15] J. Xiao, R. Zhao, S.-C. Lai, W. Jia, K.-M. Lam, Deep progressive convolutional neural network for blind super-resolution with multiple degradations, *IEEE International Conference on Image Processing (ICIP)*, IEEE 2019, pp. 2856–2860.
- [16] Y. Ju, L. Qi, J. He, X. Dong, F. Gao, J. Dong, Mps-net: learning to recover surface normal for multispectral photometric stereo, *Neurocomputing* 375 (2020) 62–70.
- [17] H. Santo, M. Samejima, Y. Sugano, B. Shi, Y. Matsushita, Deep photometric stereo network, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, IEEE 2017, pp. 501–509.
- [18] Y. Ju, M. Jian, S. Guo, Y. Wang, Z. Huiyu, J. Dong, Incorporating lambertian priors into surface normals measurement, *IEEE Trans. Instrum. Meas.* 70 (2021) 1–13.
- [19] S. Ikehata, Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces, *Proceedings of the European Conference on Computer Vision 2018*, pp. 3–18.
- [20] Y. Ju, L. Qi, H. Zhou, J. Dong, L. Lu, Demultiplexing colored images for multispectral photometric stereo via deep neural networks, *IEEE Access* 6 (2018) 30804–30818.
- [21] G. Chen, K. Han, K.-Y.K. Wong, Ps-fcn: A exible learning framework for photometric stereo, in: *Proceedings of the European Conference on Computer Vision 2018*, pp. 3–18.
- [22] B. Shi, Z. Mo, Z. Wu, D. Duan, S. Yeung, P. Tan, A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo, *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (2019) 271–284.
- [23] F. Solomon, K. Ikeuchi, Extracting the shape and roughness of specular lobe objects using four light photometric stereo, *IEEE Trans. Pattern Anal. Mach. Intell.* 18 (1996) 449–454.
- [24] A.S. Georgiades, Incorporating the Torrance and sparrow model of reectance in uncalibrated photometric stereo, *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 2003, (pp. 816–816).
- [25] H.-S. Chung, J. Jia, Efficient photometric stereo on glossy surfaces with wide specular lobes, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE 2008, pp. 1–8.
- [26] R. Ruiters, R. Klein, Heightfield and spatially varying brdf reconstruction for materials with interreflections, *Proceedings of the Computer Graphics Forum*, volume 28, Wiley Online Library 2009, pp. 513–522.
- [27] L. Chen, Y. Zheng, B. Shi, A. Subpa-Asa, I. Sato, A microfacet-based reflectance model for photometric stereo with highly specular surfaces, *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, IEEE 2017, pp. 3181–3189.
- [28] P. Tan, L. Quan, T. Zickler, The geometry of reflectance symmetries, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (2011) 2506–2520.
- [29] T. Higo, Y. Matsushita, K. Ikeuchi, Consensus photometric stereo, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE 2010, pp. 1157–1164.
- [30] M. Chandraker, J. Bai, R. Ramamoorthi, On differential photometric reconstruction for unknown, isotropic brdfs, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (2012) 2941–2955.
- [31] J. Xiao, R. Zhao, K.-M. Lam, Bayesian sparse hierarchical model for image denoising, *Signal Process. Image Commun.* 96 (2021), 116299.
- [32] C. Wang, M. Zhang, Z. Su, Y. Wu, G. Yao, H. Wang, Learning a multi-level guided residual network for single image deraining, *Signal Process. Image Commun.* 78 (2019) 206–215.
- [33] Y. Ju, X. Dong, Y. Wang, L. Qi, J. Dong, A dual-cue network for multispectral photometric stereo, *Pattern Recogn.* 100 (2020), 107162.
- [34] H. Zhu, C. Wang, Y. Zhang, Z. Su, G. Zhao, Physical model guided deep image deraining, *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE 2020, pp. 1–6.
- [35] C. Wang, H. Zhu, W. Fan, X.-M. Wu, J. Chen, Single image rain removal using recurrent scale-guide networks, *Neurocomputing* 467 (2022) 242–255.
- [36] W. Nie, Y. Zhao, D. Song, Y. Gao, Dan: deep-attention network for 3d shape recognition, *IEEE Trans. Image Process.* 30 (2021) 4371–4383.
- [37] J. Xiao, Q. Ye, R. Zhao, K.-M. Lam, K. Wan, Self-feature learning: An efficient deep lightweight network for image super-resolution, *Proceedings of the 29th ACM International Conference on Multimedia 2021*, pp. 4408–4416.
- [38] J. Li, A. Robles-Kelly, S. You, Y. Matsushita, Learning to minify photometric stereo, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019*, pp. 7568–7576.
- [39] Y. Ju, M. Jian, J. Dong, K.-M. Lam, Learning photometric stereo via manifold-based mapping, *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, IEEE 2020, pp. 411–414.
- [40] Y. Ju, K.-M. Lam, Y. Chen, L. Qi, J. Dong, Pay attention to devils: A photometric stereo network for better details, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence 2020*, pp. 694–700.
- [41] T. Tani, T. Maehara, Neural inverse rendering for general reflectance photometric stereo, *Proceedings of the International Conference on Machine Learning*, PMLR 2018, pp. 4857–4866.
- [42] Y. Ju, J. Dong, S. Chen, Recovering surface normal and arbitrary images: a dual regression network for photometric stereo, *IEEE Trans. Image Process.* 30 (2021) 3676–3690.
- [43] Z. Yao, K. Li, Y. Fu, H. Hu, B. Shi, Gps-net: Graph-based photometric stereo network, *Proceedings of the Advances in Neural Information Processing Systems*, 33, 2020, pp. 10306–10316.
- [44] W. Nie, R. Chang, M. Ren, Y. Su, A. Liu, I-gcn: incremental graph convolution network for conversation emotion detection, *IEEE Trans. Multimed.* (2021).
- [45] W. Nie, M. Ren, A. Liu, Z. Mao, J. Nie, M-gcn: multi-branch graph convolution network for 2d image-based on 3d model retrieval, *IEEE Trans. Multimed.* 23 (2020) 1962–1976.
- [46] M.K. Johnson, E.H. Adelson, Shape estimation in natural illumination, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE 2011, pp. 2553–2560.
- [47] O. Wiles, A. Zisserman, Silnet : single-and multi-view reconstruction by learning from silhouettes, *Proceedings of the British Machine Vision Conference 2017*, pp. 99.1–99.13.
- [48] W. Jakob, Mitsuba Renderer, 2010.
- [49] N. Alldrin, T. Zickler, D. Kriegman, Photometric stereo with non-parametric and spatially-varying reflectance, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE 2008, pp. 1–8.
- [50] P. Einarsson, C.-F. Chabert, A. Jones, W.-C. Ma, B. Lamond, T. Hawkins, M. Bolas, S. Sylvan, P. Debevec, Relighting human locomotion with owed reflectance fields, *Proceedings of the Eurographics Symposium on Rendering*, 2006, (pp. 76–es).
- [51] G. Chen, K. Han, B. Shi, Y. Matsushita, K.-Y.K. Wong, Deep photometric stereo for non-lambertian surfaces, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (2020) 129–142.
- [52] Y. Ju, Y. Peng, M. Jian, F. Gao, J. Dong, Learning conditional photometric stereo with high-resolution features, *Comput. Vis. Media* 8 (2022) 105–118.
- [53] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L.-Y. Duan, A.C. Kot, Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks, *Proceedings of the International Conference on Computer Vision 2019*, pp. 8549–8558.
- [54] D.B. Goldman, B. Curless, A. Hertzmann, S.M. Seitz, Shape and spatially-varying brdfs from photometric stereo, *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (2009) 1060–1071.
- [55] K. Enomoto, M. Waechter, K.N. Kutulakos, Y. Matsushita, Photometric stereo via discrete hypothesis-and-test search, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020*, pp. 2311–2319.
- [56] K.H. Cheng, A. Kumar, Revisiting outlier rejection approach for non-lambertian photometric stereo, *IEEE Trans. Image Process.* 28 (2018) 1544–1555.
- [57] T. Simchony, R. Chellappa, M. Shao, Direct analytical methods for solving poisson equations in computer vision problems, *IEEE Trans. Pattern Anal. Mach. Intell.* 12 (1990) 435–446.