Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Article Type (Research)

# Learning Conditional Photometric Stereo with High-resolution Features

Yakun Ju[1], Yuxin Peng[2], Muwei Jian[3], Feng Gao[1], and Junyu Dong[1] ✉

**Abstract** Photometric stereo aims to reconstruct the 3D geometry, by recovering the dense surface orientation of a 3D object from multiple images with varying illuminations. Traditional methods normally adopt simplified reflectance models to make the surface orientation computable. However, the general reflectance of surfaces greatly limits their applications on real-world objects. Despite deep neural networks have been employed to handle the non-Lambertian surfaces, these methods subject to blurriness and error, especially in high-frequency regions (such as crinkles and edges), caused by the spectral bias that the neural network favors low-frequency representations hence they exhibit a bias towards smooth functions. In this paper, therefore, we propose a self-learning conditional network with multi-scale features for photometric stereo, avoiding blurry reconstruction in the above regions. Our explorations include: (1) We employ a multi-scale feature fusion architecture, which keeps high-resolution representations and deep feature extraction, simultaneously. (2) We propose an improved gradient-motivated conditionally parameterized convolutions (GM-CondConv) in our photometric stereo network and provide different combinations of convolution kernels for the diversities of surfaces. Extensive experiments on public benchmark datasets show that our calibrated photometric stereo method outperforms state-of-the-art methods.

**Keywords** Photometric stereo, Normal estimation, Deep neural networks, 3D reconstruction.

## 1 Introduction

The goal of photometric stereo is to recover the dense surface orientation of a 3D object from varying shading cues, with a fixed camera, by establishing the relationship between two-dimensional images and the object geometry [16]. The earliest photometric stereo algorithm reconstructed the surface normal based on Lambertian assumption [44]. Unfortunately, the real-world objects hardly have the property of Lambertian reflectance, and therefore robust methods are needed to deal with general objects with flexible reflectance properties [21]. Traditional photometric stereo methods mainly address this problem by treating the non-Lambertian regions as the outlier [14, 45], or adopting bidirectional reflectance distribution functions (BRDFs) to model general reflectance [11, 13]. However, these traditional models are only accurate for limited categories of materials and suffer from unstable optimization.

Recently, deep learning frameworks have shown powerful abilities in various tasks [18, 41, 42]. In particular, researchers have made efforts to learn general reflectance models through deep neural networks to solve the problem of photometric stereo. DPSN [30] first addressed the non-Lambertian photometric stereo using a deep fully-connected network, to learn the surface normal in a per-pixel manner. Later, a series of methods employed the convolutional neural networks (CNNs) to better utilize the adjacent information embedded in images, such as

1 Department of Computer Science and Technology, Ocean University of China, 266100, Qingdao. E-mail: juyakun@stu.ouc.edu.cn, gaofeng@ouc.edu.cn, dongjunyu@ouc.edu.cn

2 Wangxuan Institute of Computer Technology, Peking University, 100871, Beijing. E-mail: pengyuxin@pku.edu.cn

3 School of Computer Science and Technology, Shandong University of Finance and Economics, 250002, Jinan. E-mail: jianmuweihk@163.com

Manuscript received: 2014-12-31; accepted: 2015-01-30.

PS-FCN [4], SDPS-Net [3], Manifold-PSN [19], and IRPS [38]. However, these methods suffer from the blurriness, especially in high-frequency regions (e.g. crinkles and edges). This phenomenon is caused by the spectral bias [29], where the neural network favors low-frequency representations hence they exhibit a bias towards smooth functions. Unfortunately, these regions are always where the human visual system pay attention and consequently require to be reconstructed accurately. Existing photometric stereo networks pass the input through high-to-low resolution subnetworks that are connected in series, and then raise the resolution; these procedures cause the information lost of the estimated resolution and result in the blurry. Furthermore, existing photometric stereo networks employ the same learning strategy in all surface regions. The patterns we need to learn essentially vary from plain surfaces to high-frequency surfaces, and thus the error is produced due to the same learning strategy. Therefore, it remains urgent yet challenging to develop a robust and efficient photometric stereo method that can avoid the blurry and accurately reconstruct of objects' surface orientation.

In this paper, we propose a conditional (C) deep neural network with a high-resolution (HR) structure, called CHR-PSN, for estimating the surface normal of objects. In contrast to existing methods, our framework reduces the error and blurriness, especially in those surfaces with high-frequency details. Extensive experiments on public datasets show that the our CHR-PSN achieves state-of-the-art performance. Our contributions are:

First, inspired by the High-resolution Net [36] in human pose estimation, we employ the parallel network structure for maintaining both the deep features and high-resolution details of surface normals, for the first time. We illustrate that the high-resolution of extracted features are essential to the per-pixel surface normal estimation task, which has not been explored in the learning-based or data-driven photometric stereo.

Second, we investigate an improved gradient-motivated conditionally parameterized convolutions module (GM-CondConv) [47] in the regression stage of our network, where the frequency information of surface representations is integrated into the routing function. We illustrate that the GM-CondConV module can regress the surface normal, with high-frequency details.

## 2 Related work

The imaging model establishes the relationship between the surface normal $\boldsymbol{n} \in \mathbb{R}^3$ and visual observations $\boldsymbol{I}$ in a per-pixel manner. By introducing the general BRDFs $\rho$ of the object and illumination direction $\boldsymbol{l}$ with intensity $e$, photometric stereo recovers the surface orientation from a combination of multiple images with varying illumination directions, as follows:

$$I_j = e_j \rho\left(\boldsymbol{n}, \boldsymbol{l_j}\right) \max\left(\boldsymbol{n}^\top \boldsymbol{l_j}, 0\right) + \epsilon_j \ , \tag{1}$$

where the subscript $j$ represents the index of input, $\max\left(\boldsymbol{n}^\top \boldsymbol{l_j}, 0\right)$ accounts for attached shadows, and $\epsilon$ accounts for noise (such as inter-reflections). To extend photometric stereo to work with unknown general BRDFs $\rho$ in practice, researchers investigated different strategies. We divide them into non learning-based methods and deep leaning-based methods.

### 2.1 Non learning-based methods

Generally, traditional photometric stereo technologies aim to solve the ill-posed surface normal under unknown reflectance. Here, we briefly introduce these non learning-based photometric stereo techniques, divided as sophisticated reflectance methods and outlier rejection methods. More comprehensive surveys can be found in [10, 32]

Sophisticated reflectance methods are applied to model and approximate non-Lambertian reflectance. Along this direction, many models were proposed to fit the nonlinear analytic BRDFs, such as the bivariate functions [1, 33], the Ward reflectance model [6, 9], specular spike reflectance model [5, 48], Blinn-Phong reflectance model [39], the Torrance-Sparrow reflectance model [8], etc. However, these sophisticated reflectance methods are generally useful for limited categories of surfaces as the reflectance properties are significantly changing from materials to materials.

Outlier rejection methods treat non-Lambertian regions (such as specularity and cast shadows) as outliers that should be discarded. A range of outlier rejection based photometric stereo algorithms have been proposed such as maximum-likelihood estimation [40], low rank [14, 45], RANSAC [37], and maximum feasible subsystem [49], etc. However, these methods assume the outliers are local and sparse, hardly handling the surface with broad and soft specularity.

### 2.2 Deep learning-based methods

Inspired by the powerful fitting ability of deep neural networks, deep learning-based methods have been introduced to solving the non-Lambertian photometric stereo problem. DPSN [30] first applied a fully-connected architecture for the non-Lambertian photometric stereo in a per-pixel manner. Some works introduced the observation map, which rearranges per-pixel's observation intensity according to light direction, to recover the surface normals, such as CNN-PS [12], LMPS [23], and SPLINE-Net [50]. PS-FCN [4], SDPS [3] employed the fully-convolutional network to learn the surface normal from input patches with neighborhood embedding. IRPS [38] further proposed an unsupervised learning framework that predicts the surface normals by minimizing the loss of reconstruction images. However, existing networks pass the input through high-to-low resolution subnetworks that are connected in series, and then raise the resolution, while these approaches cause the blurry of predicted surface normals.

Recently, Attention-PSN [20] proposed an adaptive attention-weighted loss to improve the performance of various surface regions. By the self-supervised weights of detail-preserving gradient loss, the method achieves better reconstruction results on high-frequency surface regions. However, we argue that the detail-preserving gradient loss can only constrain the high-frequency of surface structure but it is useless to the accuracy of predicted normal, *i.e.*, the gradient loss dilutes the supervision of normal. Furthermore, Attention-PSN only takes adaptive loss function to improve the details but ignores the impact of unsuitable kernels and receptive fields in convolutional layers, which is the essential problem of blurry in high-frequency regions.

Besides, some other reconstruction tasks also address the frequency problem into consideration. Mildenhall *et al.* [26] proposed a method for synthesizing novel views of complex scenes by optimizing an underlying continuous volumetric scene function. This method represents high-frequency scene content, by using a positional encoding to map each input 5D coordinate into a higher dimensional space. Liu *et al.* [24] introduces a wavelet-based network to remove moiré patterns, by the fact that high-frequency features might be highlighted in wavelet subbands.

## 3 Proposed Method

In this section, we will present the details of the proposed conditional deep photometric stereo network with high-resolution features. Our goal is to improve the accuracy and remove the blurriness of surface normal estimation. The architecture of the proposed CHR-PSN is shown in Figure 1.

### 3.1 Network architecture

#### 3.1.1 Feature extraction stage

As shown in Figure 1, we first fuse the input images with its illumination direction in the feature fusion stage. For an object captured under $j$ illumination directions, we expand each direction $\boldsymbol{l_j}$ to form a 3-

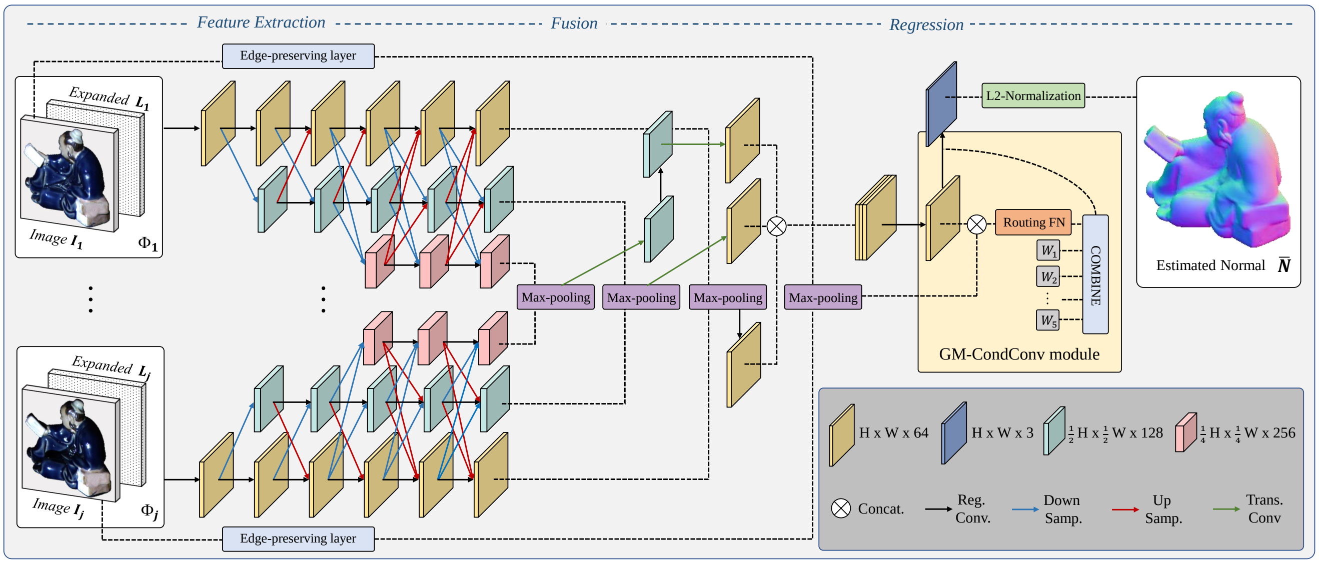

**Fig. 1** The architecture of our CHR-PSN. Reg. Conv. = regular convolution, Down Samp. = down-sampling operation, Up Samp. = up-sampling operation, and Trans. Conv = transposed convolution. We apply the Leaky-ReLU as the activation function of each layer. Our network consists of three stages, called feature extraction stage, fusion stage, and regression stage. Given an arbitrary number of images under different light directions, the feature extraction stage first extracts the multi-scale feature and representations the edge feature. Then, multi-resolution max-pooling operations are applied in the fusion stage. Finally, the regression stage with an improved GM-CondConv module infers the surface normal of the target.

channel image that has the same spatial dimension as the input image ($H \times W \times 3$), and concatenate it with its corresponding image $\boldsymbol{I_j}$ as the $\boldsymbol{\Phi_j} \in \mathbb{R}^{H \times W \times 6}$.

The feature extraction stage of our network can be seen as the $j^{th}$ multi-branch shared-weight feature extraction network, which can be expressed as:

$$\boldsymbol{\Psi_j^{FR}}, \boldsymbol{\Psi_j^{HR}}, \boldsymbol{\Psi_j^{QR}} = F_{ext}(\boldsymbol{\Phi_j}; \theta_{ext}) \ , \tag{2}$$

where $F_{ext}$ is the multi-scale feature architecture with learnable parameters $\theta_{ext}$, inspired by the High-resolution Net [36]. We employ the parallel network structure for extracting three scales feature, avoiding the feature map from low-resolution to high-resolution. Therefore, our feature extraction maintains both the deep features and high-resolution details of surface normals. As shown in Figure 1, the down-sampling operations are executed through convolutional layers with stride = 2 (double down-sampling) or 4 (twice double down-sampling), and the up-sampling operations are executed through bilinear-upsampling and $1 \times 1$ convolutional layers to adjust the channel of the feature same as high-resolution feature's channel. The fusion of high-to-low and low-to-high processes into the same-resolution features are executed through skip connections. Therefore, our feature extraction outputs three different resolution features, as full resolution (FR): $\boldsymbol{\Psi_j^{AR}} \in \mathbb{R}^{H \times W \times 64}$, half resolution (HR): $\boldsymbol{\Psi_j^{HR}} \in \mathbb{R}^{\frac{1}{2}H \times \frac{1}{2}W \times 128}$, and quarter resolution (QR): $\boldsymbol{\Psi_j^{QR}} \in \mathbb{R}^{\frac{1}{4}H \times \frac{1}{4}W \times 256}$.

Besides, we also introduce a edge-preserving layer for each $\boldsymbol{I_j}$ as:

$$\boldsymbol{\Omega_j^{FR}} = F_{edge}(\boldsymbol{I_j}) \ , \tag{3}$$

where $F_{edge}$ is the edge-preserving layer, calculated by the gradient of input image $\boldsymbol{I_j}$. $\boldsymbol{\Omega_j^{FR}} \in \mathbb{R}^{H \times W \times 3}$ is the output having high-frequency edge information, which is used in the improved CondConv module of the regression stage.

### 3.1.2 Fusion stage

In the fusion stage, we apply multi-scale max-pooling operations [4, 20] for fuse the $j$ feature to one, which makes our network can handle the arbitrary number of inputs and backpropagate the parameters. We argue that max-pooling extracts the most salient information from all features, while average-pooling may smooth out useful features and be impacted by non-activated features. Here, we choose the subscript $p$ to denote the index of position in feature as follows:

$$\boldsymbol{\Omega_{max}^{FR}} = \bigcup_{p}^{H \times W} max(\boldsymbol{\Omega_{1p}^{FR}}, \boldsymbol{\Omega_{2p}^{FR}}, \ldots, \boldsymbol{\Omega_{jp}^{FR}}) \ , \tag{4}$$

$$\boldsymbol{\Psi_{max}^{FR}} = \bigcup_{p}^{H \times W} max(\boldsymbol{\Psi_{1p}^{FR}}, \boldsymbol{\Psi_{2p}^{FR}}, \ldots, \boldsymbol{\Psi_{jp}^{FR}}) \ , \tag{5}$$

$$\boldsymbol{\Psi_{max}^{HR}} = \bigcup_{p}^{\frac{1}{2}H \times \frac{1}{2}W} max(\boldsymbol{\Psi_{1p}^{HR}}, \boldsymbol{\Psi_{2p}^{HR}}, \ldots, \boldsymbol{\Psi_{jp}^{HR}}) \ , \tag{6}$$

$$\boldsymbol{\Psi_{max}^{QR}} = \bigcup_{p}^{\frac{1}{4}H \times \frac{1}{4}W} max(\boldsymbol{\Psi_{1p}^{QR}}, \boldsymbol{\Psi_{2p}^{QR}}, \ldots, \boldsymbol{\Psi_{jp}^{QR}}) \ , \tag{7}$$

where $\boldsymbol{\Omega_{max}^{FR}}$, $\boldsymbol{\Psi_{max}^{FR}}$, $\boldsymbol{\Psi_{max}^{HR}}$, and $\boldsymbol{\Psi_{max}^{QR}}$ are the fused features.

### 3.1.3 Regression stage

The normal regression stage takes $\boldsymbol{\Omega_{max}^{FR}}$, $\boldsymbol{\Psi_{max}^{FR}}$, $\boldsymbol{\Psi_{max}^{HR}}$, and $\boldsymbol{\Psi_{max}^{QR}}$ as inputs and regresses the predicted surface normals $\boldsymbol{\bar{N}}$, by $F_{reg}$ with learnable parameters $\theta_{reg}$, as follows:

$$\boldsymbol{\bar{N}} = F_{reg}(\boldsymbol{\Omega_{max}^{FR}}, \boldsymbol{\Psi_{max}^{FR}}, \boldsymbol{\Psi_{max}^{HR}}, \boldsymbol{\Psi_{max}^{QR}}; \theta_{ext}) \ , \tag{8}$$

In the regress stage, we first employ the transposed convolution operations to up-sample the low-resolution feature $\boldsymbol{\Psi_{max}^{HR}}$ and $\boldsymbol{\Psi_{max}^{QR}}$ to the full resolution of $H \times W$ (twice transposed convolution and once regular convolution operations for $\boldsymbol{\Psi_{max}^{QR}}$, once transposed convolution operation for $\boldsymbol{\Psi_{max}^{HR}}$). As shown in Figure 1, we here employ the concatenation operation to fuse the two up-sampled features and the full resolution feature, instead of using the skip connections in the feature extraction stage.

For better reconstructing the details of objects and removing the blurry in high-frequency regions, we propose an improved GM-CondConv module in the regression stage [47], with the motivation that previous methods put the same learning strategy in all of the surface regions caused the blurry and error. By parameterizing the convolutional kernel conditionally on the input, the network can predict fine estimation for both plain surface regions and high-frequency surface regions (crinkles, edges). Particularly, we concatenate the high-frequency edge information $\boldsymbol{\Omega_{max}^{FR}}$ with the previous layer feature $\boldsymbol{x}$. We argue that the frequency information is beneficial to the classification of each learned kernel, which is better used in predict different surface normal regions. Therefore, the convolutional kernels in our GM-CondConv are parameterized as:

$$\begin{gathered} \text{GM-CondConv}(\boldsymbol{x}, \boldsymbol{\Omega_{max}^{FR}}) = \\ \sigma\left((\boldsymbol{\alpha_1} \cdot \boldsymbol{W_1} + \ldots + \boldsymbol{\alpha_n} \cdot \boldsymbol{W_n}) * (\boldsymbol{x}, \boldsymbol{\Omega_{max}^{FR}})\right) \ , \end{gathered} \tag{9}$$

where each $\boldsymbol{\alpha_i} = r_i(x, \boldsymbol{\Omega_{max}^{FR}})$ is an example-dependent scalar weight computed using a routing function with learned parameters, $n$ is the number of weights ($n = 5$ in our default setting), and $\sigma$ is the Leaky-ReLu activation function. Similar as CondConv [47], we compute the example-dependent routing weights $\boldsymbol{\alpha_i} = r_i(x, \boldsymbol{\Omega^{FR}_{max}})$ from the layer input in three steps: global average pooling, fully-connected layer, and Sigmoid activation, as:

$$\begin{aligned} &r(\boldsymbol{x}, \boldsymbol{\Omega^{FR}_{max}}) = \\ &\text{Sigmoid}(\text{GlobalAveragePool}(\boldsymbol{x}, \boldsymbol{\Omega^{FR}_{max}})\ R)\ , \end{aligned} \tag{10}$$

where $R$ is a matrix of learned routing weights mapping the pooled inputs to n expert weights. We eventually employ an $L2$ normalization that makes prediction $\bar{\boldsymbol{N}}$ be unitized.

### 3.2 Loss function and training procedure

The learning of our network is supervised by the angular error between the estimated and the ground-truth surface normals. We optimize the networks parameters $\theta_{ext}$ and $\theta_{reg}$ by minimizing the cosine similarity loss function as:

$$\mathcal{L}_{normal} = \frac{1}{HW} \sum_{p}^{HW} \left(1 - \bar{\boldsymbol{N}}_p \cdot \boldsymbol{N}_p\right)\ , \tag{11}$$

where $\bar{\boldsymbol{N}}_p$ and $\boldsymbol{N}_p$ denote the estimated normal and the ground-truth, respectively, at pixel $p$. If the estimated normal $\bar{\boldsymbol{N}}_p$ at pixel $p$ has a similar orientation as the ground-truth $\boldsymbol{N}_p$, then the $\bar{\boldsymbol{N}}_p \cdot \boldsymbol{N}_p$ will be close to 1 and the loss $\mathcal{L}_{normal}$ will approach 0.

Our network is implemented in PyTorch [28] on a RTX 2080Ti GPU, and the Adam optimizer [22] is used with default settings, where the learning rate is initially set to 0.001 and divided by 2 every 5 epochs. We train the model using a batch size of 32 for 40 epochs, with the $j = 32$ for each sample in training, whereas our network can accept an arbitrary number of $j$ in testing. Also, we set the resolution $H$ and $W = 32$ in training, and an arbitrary resolution in testing.

### 3.3 Datasets

#### 3.3.1 Training and validation datasets

We adopt two public synthetic blobby shape [17] and sculpture shape datasets [43] to train our network. Following the setup in PS-FCN [4], we render these two shape datasets with the MERL dataset [25], which contains 100 different BRDFs of real-world materials, by the physically-based raytracer Mitsuba [15]. The resolution of them is 128 × 128. Image patches of size 32 × 32 are randomly cropped for data augmentation. Eventually, we get 85212 samples in total, each sample contains 64 images with different illumination directions (random illumination directions in a space of upper semisphere). We split the samples in the dataset into the training set (84360 samples) and the validation set (852 samples).

#### 3.3.2 Testing datasets

We apply public non-Lambertian photometric stereo datasets, for evaluating our method. First, we employ the DiLiGenT benchmark dataset [32] that contains 10 objects of various shapes with complex materials. For each objects, the dataset provides 96 images under different illumination directions, with the resolution of 612 × 512. Then, we employ the Light Stage Data Gallery dataset [7] that contains six complex objects with larger resolution. Each object has up to 253 images under different illumination directions. Note that this dataset is without the ground-truth surface normal. Therefore we qualitatively evaluate our method on it.

## 4 Experimental Results

We present experiments and analysis in this section. To verify the quantitative performance of our method, we employ some widely used metrics to measure accuracy. We adopt the mean angular error (MAE) in degree to evaluate the performance of estimated surface normal, as follows:

$$\text{MAE} = \frac{1}{HW} \sum_{p}^{H \times W} \left(arccos(\bar{\boldsymbol{N}}_p \cdot \boldsymbol{N}_p)\right)\ . \tag{12}$$

We also measure the percentage (%) that the pixels with angular error less than 20°, which is denoted by $<err_{20^\circ}$. $<err_{20^\circ}$ is a metric that better measure high-frequency error terms, because the normal error in high-frequency regions are bigger.

### 4.1 Ablation Experiments and Network Analysis

We take quantitative ablation experiments on the validation. For the validation set, we report the average MAE of 852 samples (tested with 32 images).

As shown in Table 1, we summarize the results of ablation experiments. Our default method is marked as D0, with full resolution features + $\frac{1}{2}$ resolution features + $\frac{1}{4}$ resolution features in the high-resolution feature extraction stage [36], as well as the fusion of high-frequency edge information $\boldsymbol{\Omega^{FR}_{max}}$ and weights number = 5 in the GM-CondConv module of the regression stage. We first evaluate the effectiveness of multi-scale features (Experiments with IDs D0, M1, M2, M3, and M4), where the different combinations of resolution features are employed. For M1, M2, M3, and M4, we adjust the architecture of the feature extraction network, the corresponding multi-scale max-pooling fusion, and the number of concatenation in the regression stage, but maintain the GM-CondConv

**Tab. 1** MAE and $< err_{20^\circ}$ results comparison with respect to different components of the CHR-PSN on the validation set (with 32 input images).

| ID | Variants | MAE | $< err_{20^\circ}$ |
|---|---|---|---|
| D0 | Our default settings | 11.91 | **85.38%** |
| M1 | Full Resolution | 12.15 | 83.49% |
| M2 | Full Resolution + $\frac{1}{2}$ Resolution | 11.97 | 84.13% |
| M3 | $\frac{1}{2}$ Resolution + $\frac{1}{4}$ Resolution | 12.69 | 80.83% |
| M4 | Full Resolution + $\frac{1}{2}$ Resolution + $\frac{1}{4}$ Resolution+ $\frac{1}{8}$ Resolution | **11.90** | 85.35% |
| C5 | Without $\boldsymbol{\Omega_{max}^{FR}}$ | 11.99 | 84.65% |
| C6 | Weights number = 1 (Regular Conv) | 12.02 | 84.79% |
| C7 | Weights number = 3 | 11.95 | 85.05% |
| C8 | Weights number = 7 | 11.92 | 85.21% |
| L9 | Element add | 14.52 | 75.30% |

module unchanged. Note that the $\frac{1}{8}$ resolution feature in M4 has the dimensions $\in \mathbb{R}^{\frac{1}{8}H\times\frac{1}{8}W\times512}$. For the network without full resolution feature, we take down sampling at the beginning. We then evaluate the effectiveness of the improved GM-CondConv module (Experiments with IDs D0, C5, C6, C7, and C8). We test the impact of whether it fuses the edge information, and the number of weights of routing function in the GM-CondConv module. For C5, C6, C7, and C8, we only adjust the GM-CondConv module but maintain the architecture of high-resolution network unchanged. Finally, we evaluate the different fusion methods of illumination direction (Experiments with IDs D0 and L9). Our default setting uses a concatenation operation to fuse the input images and illumination directions. For L9, we test the performance of adding the value of each element instead of concatenation operation, in this case, we adjust the input channel of the first convolutional layer from 6 to 3, in the feature extraction stage.

### 4.1.1 Effectiveness of different multi-scale features

Experiments with IDs D0, M1, M2, M3, and M4 compare the performance of different combinations of feature resolution. Note that M1 has only full resolution feature, which can be seen as a fully convolutional network without up and down sampling. It can be seen that multi-scale resolution features are beneficial to the accuracy of prediction. Especially, when the network has not full resolution features (M3), the performance is obviously worse. It illustrates that the resolution of features has a crucial impact on the performance of the model in the per-pixel surface normal recovery task. Unfortunately, previous deep learning-based photometric stereo methods belong to the classification of M3 (without the branch of high-resolution feature). Also, compared D0 with M1, M2 and M4, we can see that the deep features improve the performance of prediction to some extent. However, the improvement is quite slight after adding $\frac{1}{8}$ resolution feature, and the $\frac{1}{8}$ resolution feature significantly increases the parameters and training time. This might be because such deep feature contains less detailed information but high-level semantic information, which is useless for the per-pixel prediction task. Therefore, we select full resolution features + $\frac{1}{2}$ resolution features + $\frac{1}{4}$ resolution features in the high-resolution feature extraction stage [36].

### 4.1.2 Effectiveness of fusing high-frequency information in routing function

Experiment with IDs D0 and C5 show the influence of fusing high-frequency edge information $\boldsymbol{\Omega_{max}^{FR}}$ in the routing function of GM-CondConv module. We can see that the angular error and $< err_{20^\circ}$ of the validation set are lower when the edge information is taken into account. This might be explained by the fact that the improved routing function will take the high-frequency information into the self-learned weights, which is beneficial to the GM-CondConv module for estimating different frequency surface regions (such as crinkle and plain).We also show a visualized sample of "Buddha" in Fig. 2. The comparisons between the CondConv (ID = C5) and GM-CondConv (ID = D0) show that using GM-CondConv will further improve the performance of high-frequency areas.

### 4.1.3 Effectiveness of weights number in GM-CondConv module

Referring to the experiments with IDs D0, C6, C7, our method increased with the number of weights in GM-CondConv. Note that weights number = 1 means there are only one convolution kernel and no dynamic weight. These comparisons show the effectiveness of our improved GM-CondConv module. Also, compared with default settings, more weights in GM-CondConv can not improve the accuracy continuously. Our method performs better when the number of the weights is 5, according to the above experiments.

### 4.1.4 Effectiveness of illumination direction fusion methods

Experiment with IDs D0 L9 show the influence of different fusing methods. We can see that the angular error and $< err_{20^\circ}$ of the validation set

**Tab. 2** Comparison of different methods on the DiLiGenT benchmark dataset. All methods are evaluated with 96 images. Here, we measure MAE in degrees.

| Method | Ball | Bear | Buddha | Cat | Cow | Goblet | Harvest | Pot1 | Pot2 | Reading | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 4.10 | 8.39 | 14.92 | 8.41 | 25.60 | 18.50 | 30.62 | 8.89 | 14.65 | 19.80 | 15.39 |
| Matrix rank = 3 | 2.54 | 7.32 | 11.11 | 7.21 | 25.70 | 16.25 | 29.26 | 7.74 | 14.09 | 16.17 | 13.74 |
| Rank minimization | 2.06 | 6.50 | 10.91 | 6.73 | 25.89 | 15.70 | 30.01 | 7.18 | 13.12 | 15.39 | 13.35 |
| Multi-Ward models | 3.21 | 6.62 | 14.85 | 8.22 | 9.55 | 14.22 | 27.84 | 8.53 | 7.90 | 19.07 | 12.00 |
| Bivariate BRDF | 3.34 | 7.11 | 10.47 | 6.74 | 13.05 | 9.71 | 25.95 | 6.64 | 8.77 | 14.19 | 10.60 |
| Bi-polynomial | 1.74 | 6.12 | 10.60 | 6.12 | 13.93 | 10.09 | 25.44 | 6.51 | 8.78 | 13.63 | 10.30 |
| DPSN | 2.02 | 6.31 | 12.68 | 6.54 | 8.01 | 11.28 | 16.86 | 7.05 | 7.86 | 15.51 | 9.41 |
| IRPS | **1.47** | 5.79 | 10.36 | **5.44** | 6.32 | 11.47 | 22.59 | **6.09** | 7.76 | **11.03** | 8.83 |
| PS-FCN | 2.82 | 7.55 | 7.91 | 6.16 | 7.33 | 8.60 | 15.85 | 7.13 | 7.25 | 13.33 | 8.39 |
| Attention-PSN | 2.93 | **4.86** | 7.75 | 6.14 | 6.86 | 8.42 | 15.44 | 6.92 | 6.97 | 12.90 | 7.92 |
| CHR-PSN (Ours) | 2.26 | 6.35 | **7.15** | 5.97 | **6.05** | **8.32** | **15.32** | 7.04 | **6.76** | 12.52 | **7.77** |

are best when using concatenation operation (our default settings). Also, the performance of prediction is severely decreased when using the add operation between the input image and the illumination direction. We argue that the network can hardly decouple the feature that is numerically added between image and illumination.

## 4.2 Evaluation on the DiLiGenT benchmark dataset

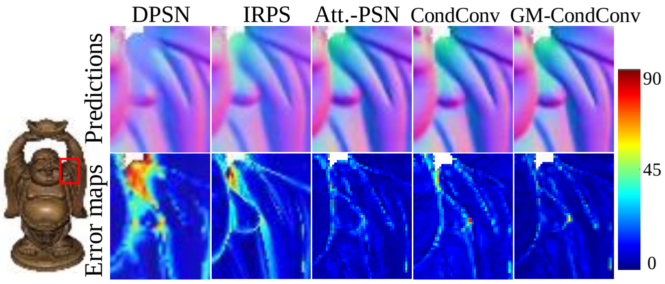

**Fig. 2** An enlarged sample "Buddha" from DiLiGenT dataset [32]. Att.-PSN is short for Attention-PSN. CondConv represents using original CondConv module [47] (ID = C5 in Tab. 1), while GM-CondConv represents our default model.

We compare our method with both non learning-based methods and recently deep learning-based methods in terms of achievable MAE, on the DiLiGenT benchmark dataset [32]. For non learning-based methods, we evaluate the Least squares (baseline) method [44], Rank minimization[45] and Matrix rank = 3 [14] of outlier rejection method. We also evaluate the sophisticated reflectance methods, such as Multi-Ward models [9], Bivariate BRDF [13], and Bi-polynomial [34]. For deep learning-based methods, we compared our method with DPSN [30], IRPS [38], PS-FCN [4], and Attention-PSN [20] in 96 input images. Quantitative results are reported on Table 2. Figure 3 presents the visualized results with the top four accuracy deep learning-based photometric stereo methods, including Attention-PSN [20], PS-FCN [4], IRPS [38], and DPSN [30], as well as the Baseline method (Least square) [44]. As shown in Figure 3, we illustrate the performance of our method on high-frequency regions, such as the face of "Buddha" and the flower of "Pot2", and cast shadows regions, such as the shoulder of "Buddha" and the base of "Goblet". It can be seen that our method is more accurate on those regions with cast shadows and crinkles.

We also show an enlarged sample of "Buddha" in Fig. 2, with details. We can see that the last three comparisons, which take the high-frequency information into consideration, achieve much better accuracy on crinkles and edges. Specifically, our default settings (using improved GM-CondConv) recover less error in high-frequency areas, compared using CondConv module (without high-frequency information $\boldsymbol{\Omega}^{\boldsymbol{F}\,\boldsymbol{R}}_{\boldsymbol{max}}$ in routing function, ID = C5 in Tab.1).

### 4.2.2 Discussion on limitations

We notice that the proposed CHR-PSN does not achieve the best performance on some objects, such as "Ball"and "Bear". We also illustrate some failed cases in Figure 4. In these objects, our method only obtains

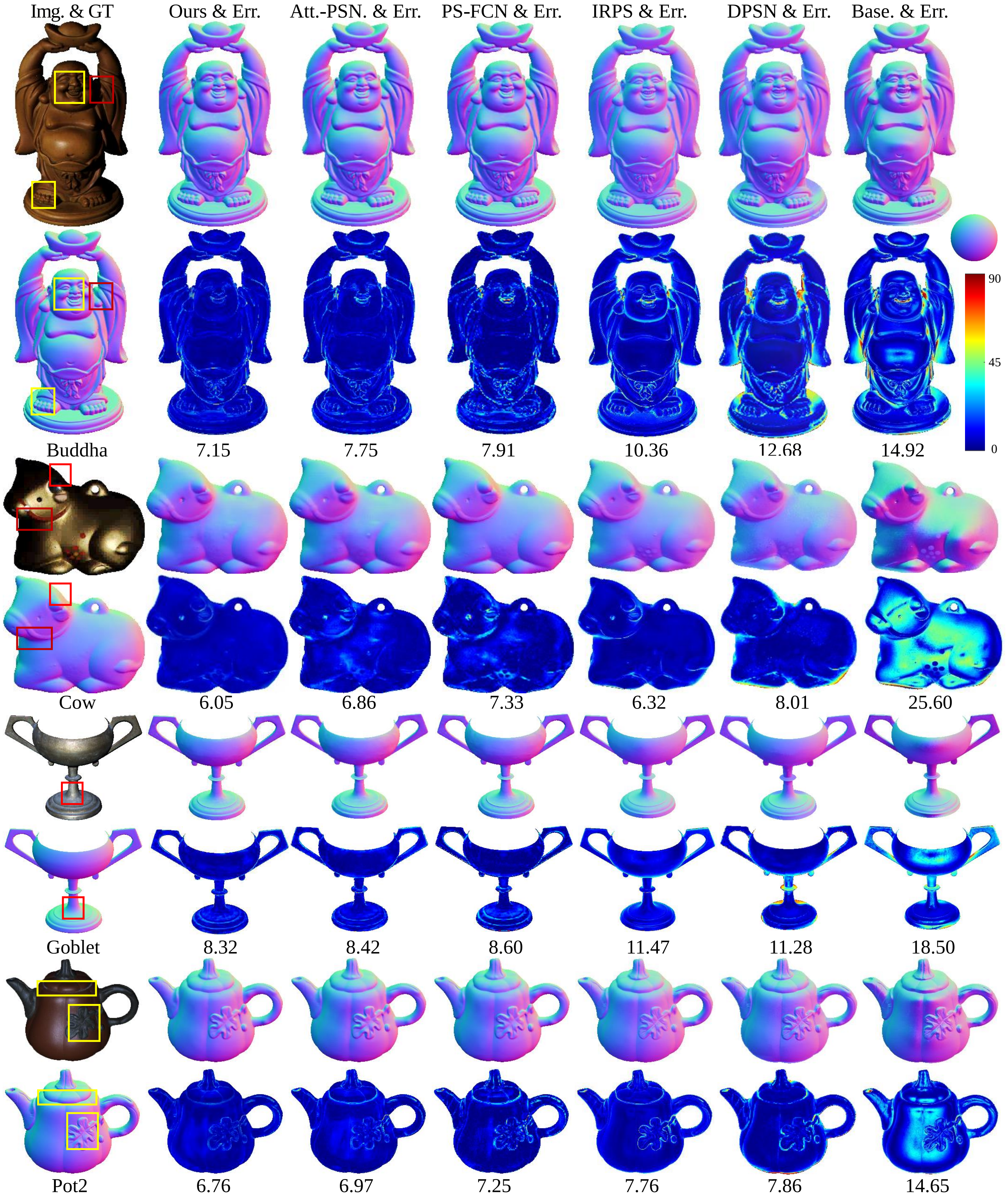

**Fig. 3** Comparisons on the Buddha, Cow, Goblet, and Pot2 of DiLiGenT benchmark dataset. The yellow boxes in the observed images and ground-truth surface normals are regions with high-frequency surface (such as crinkles), while the red boxes are regions with cast shadows. Att.-PSN is short for Attention-PSN [20]. Base. is short for Baseline least square method [44] The contrast of observations is adjusted for easy viewing.

sub-optimal performance. We argue that the objects like Ball and Bear have the plain surface normal and approximate Lambertian reflectance. In these cases, we argue that the high-resolution feature extraction of our method and GM-CondConv module are excess. IRPS [38] performs very well on these objects because it introduces the reconstruction loss to learn the surface normal, where an approximate Lambertian surface and simple structure is beneficial to the inverse rendering. However, we can see that our method still outperform Attention-PSN and IRPS on non-Lambertian regions

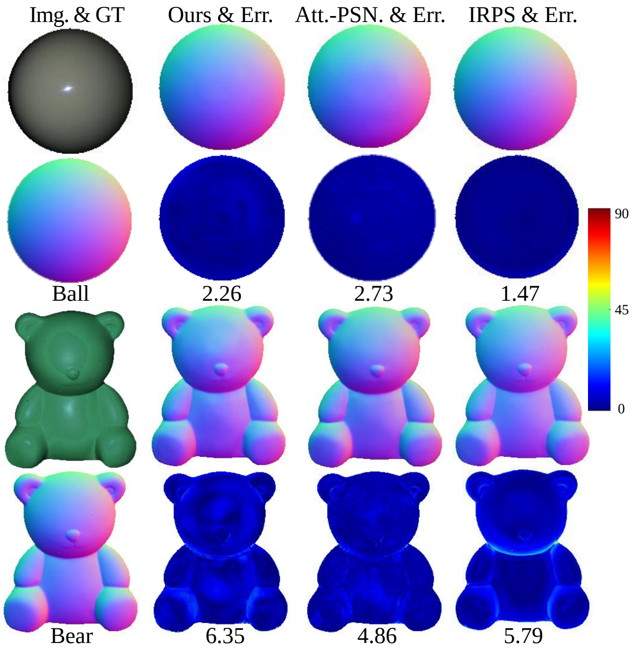

**Fig. 4** Quantitative results on Ball and Bear from the DiLiGenT benchmark dataset. The contrast of observations is adjusted for easy viewing.

#### 4.2.3 Evaluation on the less number of input images

We further evaluated our method against several methods with sparse inputs (with 10 input images). Our method employ the max-pooling operation to handle arbitrary input number of image, which is of piratical use. For non learning-based methods, we evaluate the Least squares (baseline) method [44], Bi-polynomial [34], and Matrix rank = 3 [14]. For deep learning-based method, we evaluate the CNN-PS [12], SPLINE-Net [50], LMPS [23], and PS-FCN [4]. We summarize the comparisons in Table 3.

It can be seen that our method outperforms others on average MAE of the DiLiGenT dataset and achieves state-of-the-art accuracy on most objects. We also visualize the average MAE of the DiLiGenT dataset from sparse input (8) to dense input (96), as shown in Figure 5. We compare our method with PS-FCN [4], which also uses the max-pooling operation to handle the different numbers of input images with once training. We can see that our method outperforms PS-FCN on all numbers of input image (Both methods are trained with 32 input images).

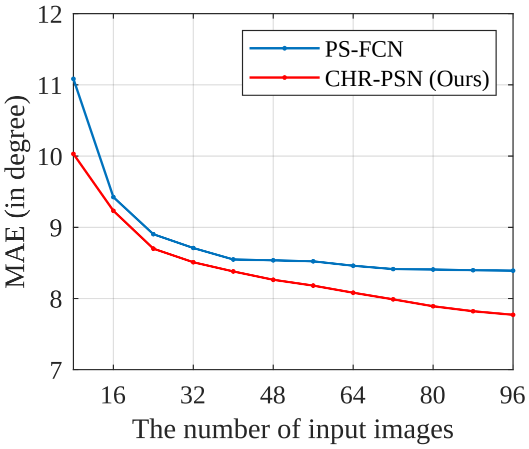

**Fig. 5** The comparison of MAE under different numbers of input images.

#### 4.2.4 Extension on uncalibrated photometric stereo

Furthermore, we report the superior performance of our method on uncalibrated conditions. In actual applications, there are conditions where the directions of illuminations $\boldsymbol{l_j}$ are unknown. Our method can be easily extended to handle uncalibrated photometric stereo by removing the illumination direction from the input (as the $\boldsymbol{\Phi_j} \in \mathbb{R}^{H\times W\times 3}$, which only includes the RGB-channel image). To verify the potential of our method, we train the model without illumination direction (aslo use 32 images for one sample) and test it on the DiLiGenT benchmark dataset [32] with 96 images. The results are reported in Table 4. We compare our method (uncalibrated) with several uncalibrated photometric stereo methods, such as entropy minimization [2], self-calibrating [31], reflectance symmetry [46], diffuse maxima [27], and UPS-FCN (for uncalibrated)[3]. Our method (uncalibrated) outperformed existing methods in terms of the average MAE, except SDPS-Net [3]. SDPS-Net is specially designed for uncalibrated condition (learn the illumination direction solely), while our method can be both used in the calibrated and uncalibrated conditions, which is not designed for calibrated condition.

**Tab. 3** Comparison of different methods on the DiLiGenT benchmark dataset. We note that all methods are evaluated with 10 images for MAE in degrees.

| Method | Ball | Bear | Buddha | Cat | Cow | Goblet | Harvest | Pot1 | Pot2 | Reading | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 5.09 | 11.59 | 16.25 | 9.66 | 27.90 | 19.97 | 33.41 | 11.32 | 18.03 | 19.86 | 17.31 |
| Bi-polynomial | 5.24 | 9.39 | 15.79 | 9.34 | 26.08 | 19.71 | 30.85 | 9.76 | 15.57 | 20.08 | 16.18 |
| Matrix rank = 3 | **3.33** | 7.62 | 13.36 | 8.13 | 25.01 | 18.01 | 29.37 | 8.73 | 14.60 | 16.63 | 14.48 |
| CNN-PS | 9.11 | 14.08 | 14.58 | 11.71 | 14.04 | 15.48 | 19.56 | 13.23 | 14.65 | 16.99 | 14.34 |
| PS-FCN | 4.02 | 7.18 | 9.79 | 8.80 | 10.51 | 11.58 | 18.70 | 10.14 | 9.85 | 15.03 | 10.51 |
| SPLINE-Net | 4.96 | **5.99** | 10.07 | 7.52 | 8.80 | 10.43 | 19.05 | 8.77 | 11.79 | 16.13 | 10.35 |
| LMPS | 3.97 | 8.73 | 11.36 | **6.69** | 10.19 | 10.46 | 17.33 | **7.30** | 9.74 | 14.37 | 10.02 |
| CHR-PSN (Ours) | 3.91 | 7.84 | **9.59** | 8.10 | **8.54** | **10.36** | **17.21** | 9.65 | **9.61** | **14.35** | **9.92** |

**Tab. 4** Comparison of results for uncalibrated photometric stereo on the DiLiGenT benchmark dataset. All the methods are evaluated with 96 images for MAE in degrees.

| Method | Ball | Bear | Buddha | Cat | Cow | Goblet | Harvest | Pot1 | Pot2 | Reading | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Entropy minimization | 7.27 | 16.81 | 32.81 | 31.45 | 54.72 | 46.54 | 61.70 | 18.37 | 49.16 | 53.65 | 37.25 |
| Self-calibrating | 8.90 | 11.98 | 15.54 | 19.84 | 22.73 | 48.79 | 73.86 | 16.68 | 50.68 | 26.93 | 29.59 |
| Reflectance symmetry | **4.39** | **6.42** | 13.19 | 36.55 | 19.75 | 20.57 | 55.51 | **9.39** | 14.52 | 58.96 | 23.93 |
| Diffuse maxima | 4.77 | 9.07 | 14.92 | **9.54** | 19.53 | 29.93 | 29.21 | 9.51 | 15.90 | 24.18 | 16.66 |
| UPS-FCN | 6.62 | 11.23 | 15.87 | 14.68 | **11.91** | 20.72 | 27.79 | 13.98 | 14.19 | 23.26 | 16.02 |
| Ours (Uncalibrated) | 5.61 | 10.80 | **12.48** | 13.95 | 12.44 | **17.84** | **23.39** | 13.62 | **13.79** | **20.78** | **14.47** |
| SDPS-Net | 2.77 | 6.89 | 8.97 | 8.06 | 8.48 | 11.91 | 17.43 | 8.14 | 7.50 | 14.90 | 9.51 |

### 4.3 Evaluation on the Light Stage Data Gallery dataset

We further qualitatively evaluated our method on a more complex dataset with general non-Lambertian materials. Figure 6 shows the results of our method (test with random 150 of 253 total images) on objects Kneeling, Helmet, and Standing. We show the qualitative outcomes in this experiment, due to the absence of ground-truth surface normals. Owing to the memory limit of GPU, we test the Light Stage Data Gallery with 64 input images (calibrated illumination directions).

As shown in Fig. 6, the estimated normal keeps the details without blur, such as the hair of the Kneeling, and the screws of the Helmet. The predicted surface normal and 3D reconstruction convincingly reflect the shapes of the objects, with accurate detail. Besides, the belt of the Kneeling illustrates our performance on cast shadows. However, we also notice that the predicted surface normal of the object Kneeling meets some blurry and noise. We argue that the poor quality of the observations of Kneeling with noise, where the high-frequency noise existing in observation may affect the GM-CondConv module of our method.

## 5 Conclusions

In this paper, we have proposed a conditional photometric stereo network with high-resolution feature extraction architecture. Compared with previous deep learning approaches regress the surface normals from the down-sampled feature map, we employ the multi-scale parallel architecture which enhances the details of prediction. Furthermore, we employ an improved GM-ConvCond module in the regression stage which considers the frequency of surfaces. Therefore, our method outperforms others in high-frequency regions such as crinkles and edges. Ablation experiments have illustrated that our method performs more accurate reconstruction. Extensive quantitative and qualitative comparisons on the DiLiGenT benchmark and the Light Stage Data Gallery have shown that our method outperforms the state-of-the-art methods. Despite offering state-of-the-art performance, our method can be further improved. Firstly, our method only achieves sub-

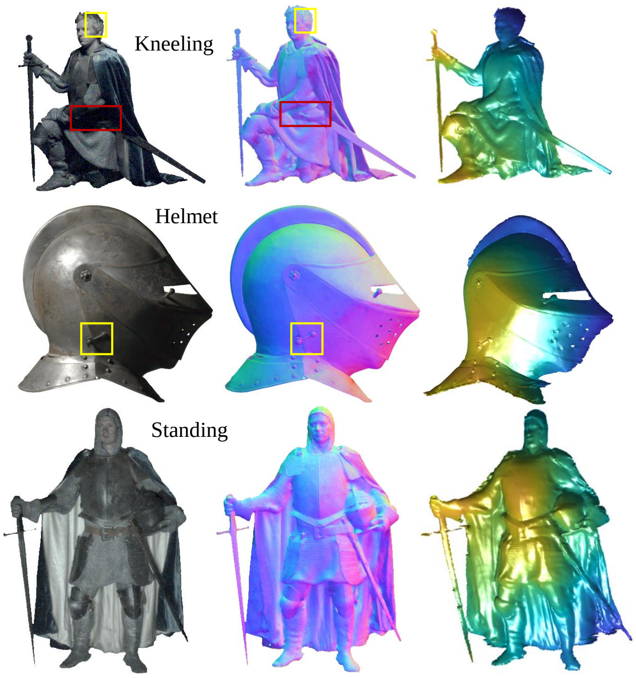

**Fig. 6** Qualitative results of our method on objects Kneeling, Helmet, and Standing. The yellow boxes in the observed images and ground-truth surface normals are regions with high-frequency surface (such as crinkles), while the red boxes are regions with cast shadows. The contrast of observations is adjusted for easy viewing. The 3D reconstructions after predicted surface normals are recovered by [35].

optimal performance on some objects with very simple structure, where the high-resolution feature extraction and GM-CondConv are excess in these cases. Secondly, the training time of our method is longer than other deep learning-based photometric stereo, which is due to our much bigger network architecture. In the future, we will further design the architecture of feature extractor to better and fast predict the surface normal.

## Acknowledgements

The work was supported by the National Key Scientific Instrument and Equipment Development Projects of China (41927805), the National Natural Science Foundation of China (61501417, 61976123), the Key Development Program for Basic Research of Shandong Province (ZR2020ZD44), and the Taishan Young Scholars Program of Shandong Province.

## References

[1] N. Alldrin, T. Zickler, and D. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[2] N. G. Alldrin, S. P. Mallick, and D. J. Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–7, 2007.

[3] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8747, 2019.

[4] G. Chen, K. Han, and K.-Y. K. Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[5] T. Chen, M. Goesele, and H.-P. Seidel. Mesostructure from specularity. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1825–1832. IEEE, 2006.

[6] H.-S. Chung and J. Jia. Efficient photometric stereo on glossy surfaces with wide specular lobes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[7] P. Einarsson, C.-F. Chabert, A. Jones, W.-C. Ma, B. Lamond, T. Hawkins, M. Bolas, S. Sylwan, and P. Debevec. Relighting human locomotion with flowed reflectance fields. In *Proceedings of the 17th Eurographics conference on Rendering Techniques*, pages 183–194, 2006.

[8] A. S. Georghiades. Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 816–823. IEEE, 2003.

[9] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and spatially-varying brdfs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1060–1071, 2009.

[10] S. Herbort and C. Wöhler. An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods. *3D Research*, 2(3):4, 2011.

[11] T. Higo, Y. Matsushita, and K. Ikeuchi. Consensus photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1157–1164. IEEE, 2010.

[12] S. Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proceedings of the*

*European Conference on Computer Vision*, pages 3–18, 2018.

[13] S. Ikehata and K. Aizawa. Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2186, 2014.

[14] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Robust photometric stereo using sparse regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 318–325. IEEE, 2012.

[15] W. Jakob. Mitsuba renderer, 2010.

[16] M. Jian, J. Dong, M. Gong, H. Yu, L. Nie, Y. Yin, and K.-M. Lam. Learning the traditional art of chinese calligraphy via three-dimensional reconstruction and assessment. *IEEE Transactions on Multimedia*, 22(4):970–979, 2019.

[17] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2553–2560. IEEE, 2011.

[18] Y. Ju, X. Dong, Y. Wang, L. Qi, and J. Dong. A dual-cue network for multispectral photometric stereo. *Pattern Recognition*, 100:107162, 2020.

[19] Y. Ju, M. Jian, J. Dong, and K.-M. Lam. Learning photometric stereo via manifold-based mapping. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 411–414. IEEE, 2020.

[20] Y. Ju, K. Lam, Y. Chen, L. Qi, and J. Dong. Pay attention to devils: A photometric stereo network for better details. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 694–700, 2020.

[21] M. Khanian, A. S. Boroujerdi, and M. Breuß. Photometric stereo for strong specular highlights. *Computational Visual Media*, 4(1):83–102, 2018.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[23] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita. Learning to minify photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7568–7576, 2019.

[24] L. Liu, J. Liu, S. Yuan, G. Slabaugh, A. Leonardis, W. Zhou, and Q. Tian. Wavelet-based dual-branch network for image demoiréing. In *European Conference on Computer Vision*. Springer, 2020.

[25] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Transactions on Graphics*, 22(3):759–769, 2003.

[26] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

[27] T. Papadhimitri and P. Favaro. A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima. *International journal of computer vision*, 107(2):139–154, 2014.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in neural information processing systems*, pages 8026–8037, 2019.

[29] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[30] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita. Deep photometric stereo network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 501–509, 2017.

[31] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan. Self-calibrating photometric stereo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1118–1125, 2010.

[32] B. Shi, Z. Mo, Z. Wu, D. Duan, S. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):271–284, 2019.

[33] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1078–1091, 2013.

[34] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):1078–1091, 2014.

[35] T. Simchony, R. Chellappa, and M. Shao. Direct analytical methods for solving poisson equations in computer vision problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):435–446, 1990.

[36] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[37] K. Sunkavalli, T. Zickler, and H. Pfister. Visibility subspaces: Uncalibrated photometric stereo with shadows. In *Proceedings of the European Conference on Computer Vision*, pages 251–264. Springer, 2010.

[38] T. Taniai and T. Maehara. Neural inverse rendering for general reflectance photometric stereo. In *Proceedings of the International Conference on Machine Learning*, pages 4857–4866, 2018.

[39] S. Tozza, R. Mecca, M. Duocastella, and A. Del Bue.

Direct differential photometric stereo shape recovery of diffuse and specular surfaces. *Journal of Mathematical Imaging and Vision*, 56(1):57–76, 2016.

[40] F. Verbiest and L. Van Gool. Photometric stereo with coherent outlier handling and confidence estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[41] K. Wei, C. Deng, and X. Yang. Lifelong zero-shot learning. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization.

[42] K. Wei, M. Yang, H. Wang, C. Deng, and X. Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3741–3749, 2019.

[43] O. Wiles and A. Zisserman. Silnet: Single-and multi-view reconstruction by learning from silhouettes. In *Proceedings of the British Machine Vision Conference*, 2017.

[44] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[45] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Proceedings of the Asian Conference on Computer Vision*, pages 703–717. Springer, 2010.

[46] Z. Wu and P. Tan. Calibrating photometric stereo by holistic reflectance symmetry analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1505, 2013.

[47] B. Yang, G. Bender, Q. V. Le, and J. Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1307–1318, 2019.

[48] S.-K. Yeung, T.-P. Wu, C.-K. Tang, T. F. Chan, and S. J. Osher. Normal estimation of a transparent object using a video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4):890–897, 2014.

[49] C. Yu, Y. Seo, and S. W. Lee. Photometric stereo from maximum feasible lambertian reflections. In *Proceedings of the European Conference on Computer Vision*, pages 115–126, 2010.

[50] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L.-Y. Duan, and A. C. Kot. Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8549–8558, 2019.

**Yakun Ju** Yakun Ju received the B.Sc degree from Sichuan University, Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree in computer application technology with the Department of Computer Science and Technology, Ocean University of China, Qingdao, China, supervised by professor Junyu Dong. His research interests include 3D reconstruction, deep learning, and image processing.

**Yuxin Peng** Yuxin Peng received the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2003. He is currently the Boya Distinguished Professor with the Wangxuan Institute of Computer Technology, Peking University. He has authored more than 160 articles in refereed international journals and conference proceedings, including more than 70 articles on International Journal of Computer Vision, the IEEE Transactions on Image processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Multimedia, the IEEE Transactions on Cybernetics, ACM MM, ICCV, CVPR, IJCAI, and AAAI. He has submitted 42 patent applications and been granted 24 of them. His current research interests mainly include cross-media analysis and reasoning, image and video recognition and understanding, and computer vision. Dr. Peng was a recipient of the First Prize of the Beijing Science and Technology Award in 2016 (ranking first) and the National Science Fund for Distinguished Young Scholars of China in 2019. He led his team to win the first place in video instance search evaluation of TRECVID in the recent years.

**Muwei Jian** Muwei Jian received the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, in October 2014. He was a Lecturer with the Department of Computer Science and Technology, Ocean University of China, from 2015 to 2017. He is currently a Professor and Ph.D. Supervisor with the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China. His current research interests include human face recognition, image and video processing, machine learning, and computer vision.

**Feng Gao** Feng Gao received his B.Sc. degree from the Department of Computer Science, Chongqing University, Chongqing, China in 2008, and received the Ph.D. degree from the Department of Computer Science and Engineering, Bei- hang University, Beijing, China in 2015. He is currently an associate professor in the Department of Computer Sci- ence and Technology in Ocean University of China. His research interests include computer vision and remote sensing.

**Junyu Dong** Junyu Dong received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, U.K., in 2003. He joined Ocean University of China in 2004. He is currently a Professor and the Vice-Dean of the College of Information Science and Engineering, Ocean University of China. His research interests include computer vision, underwater image processing, and machine learning, with more than ten research projects supported by the NSFC, MOST, and other funding agencies.