Deep Learning Methods for Calibrated Photometric Stereo and Beyond

Yakun Ju, *Member, IEEE,* Kin-Man Lam, *Senior Member, IEEE,* Wuyuan Xie, *Member, IEEE,* Huiyu Zhou, Junyu Dong, *Member, IEEE,* and Boxin Shi, *Senior Member, IEEE*

Abstract—Photometric stereo recovers the surface normals of an object from multiple images with varying shading cues, *i.e.*, modeling the relationship between surface orientation and intensity at each pixel. Photometric stereo prevails in superior per-pixel resolution and fine reconstruction details. However, it is a complicated problem because of the non-linear relationship caused by non-Lambertian surface reflectance. Recently, various deep learning methods have shown a powerful ability in the context of photometric stereo against non-Lambertian surfaces. This paper provides a comprehensive review of existing deep learning-based calibrated photometric stereo methods utilizing orthographic cameras and directional light sources. We first analyze these methods from different perspectives, including input processing, supervision, and network architecture. We summarize the performance of deep learning photometric stereo models on the most widely-used benchmark data set. This demonstrates the advanced performance of deep learning-based photometric stereo methods. Finally, we give suggestions and propose future research trends based on the limitations of existing models.

Index Terms—Photometric stereo, deep learning, non-Lambertian, surface normals.

1 Introduction

Acquiring three-dimensional (3D) geometry from two-dimensional (2D) scenes is a fundamental problem in computer vision. It aims to establish computational models that allow computers to perceive the external 3D world. Unlike geometric approaches (such as multi-view stereo and binocular) that use different viewpoint scenes to compute 3D points, photometric stereo [1] perceives the shape of an object from varying shading cues observed under different lighting conditions with a fixed viewpoint. Compared to geometric methods that generally reconstruct rough shapes, photometric methods can acquire more detailed local reconstruction. Therefore, photometric stereo plays a mainstream role in many high-precision surface reconstruction tasks, such as cultural relic reconstruction [2], seabed mapping [3], moon surface reconstruction [4], and industrial defect detection [5], *etc*. As shown in Fig. 1, photometric stereo methods obtain detailed shape reconstructions from multiple images under different illuminations. In this survey, we take the object "Reading" from the DiLiGenT benchmark [6] as a visual example, which has spatially varying and non-Lambertian materials with strong specularity and shadow.

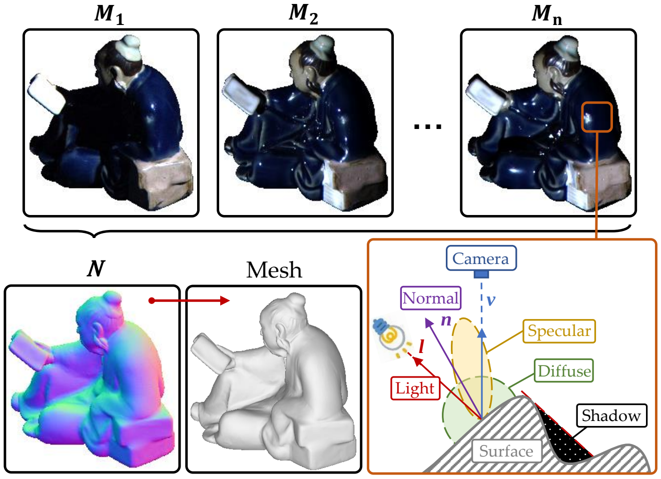


Fig. 1. The schematic of photometric stereo. The orange box shows the general surface reflectance.

Classic photometric stereo [1] assumed that only the Lambertian (diffuse) reflectance exists on the surface of the target object. Under the Lambertian assumption, the surface normal can be easily solved by the least squares method, because the reflection intensity $\boldsymbol{M}$ is linearly proportional to the angle between the normal $\boldsymbol{n}$ and incident light $\boldsymbol{l}$, as follows:

$$\boldsymbol{M} \propto \boldsymbol{l}^\top \boldsymbol{n}. \tag{1}$$

However, real-world objects barely have the property of Lambertian reflectance. The non-Lambertian property of surfaces (as shown in the orange box in Fig. 1) affects the proportional relationship of Eq. 1. Mathematically, we express the non-Lambertian property via the bidirectional

- *Yakun Ju is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: kelvin.yakun.ju@gmail.com).*
- *Kin-Man Lam is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: enkmlam@polyu.edu.hk).*
- *Wuyuan Xie is with the Research Institute for Future Media Computing, Shenzhen University, Shenzhen, China (e-mail: wuyuan.xie@gmail.com).*
- *Huiyu Zhou is with the Department of Informatics, University of Leicester, Leicester, UK (e-mail: hz143@leicester.ac.uk).*
- *Junyu Dong is with the Faculty of Information Science and Engineering and the Institute for Advanced Ocean Study, Ocean University of China, Qingdao (e-mail: dongjunyu@ouc.edu.cn).*
- *Boxin Shi is with the National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China (e-mail: shiboxin@pku.edu.cn).*
- *Corresponding authors: Junyu Dong and Boxin Shi.*

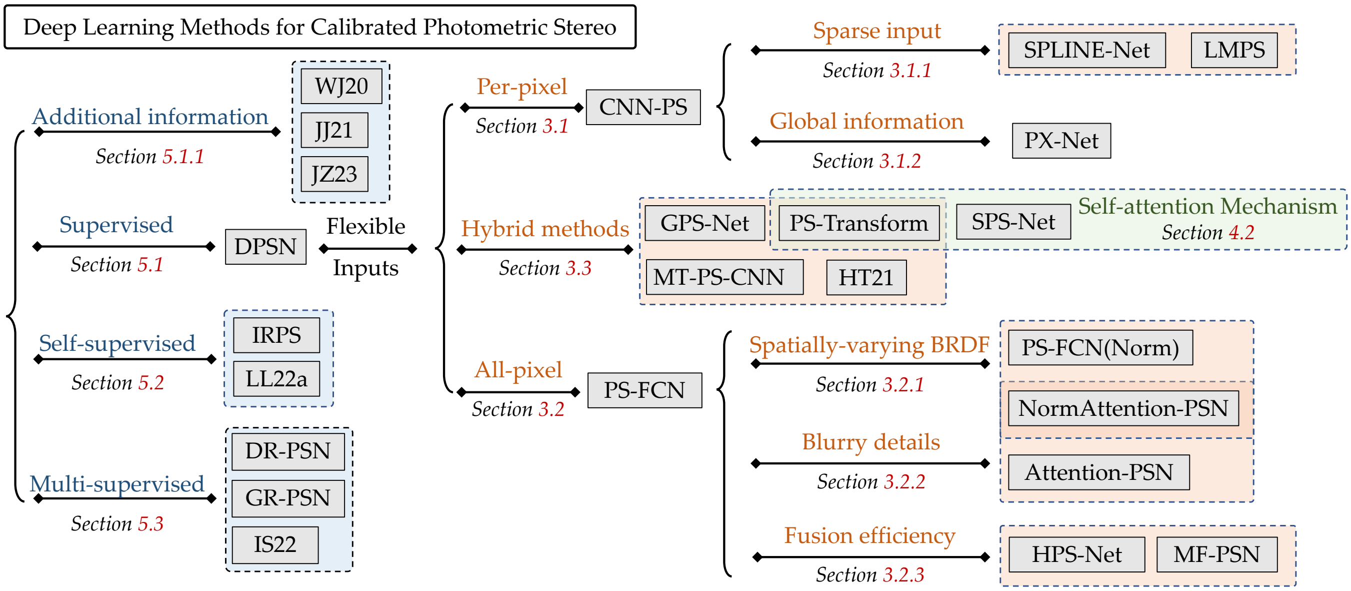


Fig. 2. Overview of the main deep learning methods for calibrated photometric stereo.

reflectance distribution function (BRDF), depending on the material of the object. According to the previous taxonomies [6], [7], [8], plenty of work has addressed non-Lambertian photometric stereo by modeling BRDF [9], [10], [11], rejecting outlier regions [12], [13], [14], or setting exemplars [15], [16]. Nevertheless, designing appropriate reflectance models using general parametric BRDFs for photometric stereo is challenging, since these non-learning models tend to be accurate only for specific materials and often involve unstable optimization processes. In this context, early nonparametric attempts based on shallow artificial neural networks were introduced to establish a mapping between complex reflectance observations and surface normal [17], [18], [19]. However, these models were restricted to limited materials that lack practical applications or require pre-training with a reference object with the same material as the target.

In 2017, DPSN [20] first attempted to use modern deep neural network architecture in the context of photometric stereo. It established the learning-based photometric stereo framework that more flexible mapping from reflectance observations to surface normal, breaking through the per-material-per-train limitation in early methods. DPSN [20] showed superior performance on non-Lambertian surfaces compared with traditional hand-crafted models which explicitly estimate the BRDF parameters and decouple the surface normals. However, this method required a fixed number and order of illumination directions during training and testing, which limited its generalization.

To enhance the generalization, various deep learning-based approaches have been introduced. This paper specifically concentrates on deep learning-based calibrated photometric stereo methods utilizing orthographic cameras and directional light sources. We categorize and summarize these methods from different perspectives, including input processing, supervision, and network architecture, the overview framework is shown in Fig. 2.

In this paper, we first categorize these deep learning-based calibrated photometric stereo methods based on how they process the input images, as per-pixel methods (*i.e.*, by the observation map operation [21] to record the intensity of each pixel) or all-pixel methods (*i.e.*, by using aggregation model [22] to fuse whole patches). Different from the recent summary [23] that only five calibrated learning-based photometric stereo models were listed, we comprehensively summarize and discuss the pros and cons of the various methods and how they evolve within these two categories. Additionally, we introduce a new classification, known as hybrid methods, which leverage both pixel- and patch-wise characteristics to enhance performance.

Second, the complexity and the number of parameters in learning-based models have significantly increased. Many advanced modules were integrated into surface-normal recovery tasks, such as ResNet [24], DenseNet [25], HR-Net [26], Transformer [27], *etc*. Similarly, the selection of synthetic training data sets for photometric stereo became more diverse, *i.e.*, rendering with MERL BRDF [28] or Disney's principled BSDF [29]. In this paper, we also conduct a comprehensive summary and discussion of the network architectures and data sets utilized in previous deep learning-based photometric stereo methods.

In addition, we analyze the literature from the perspective of supervision, *i.e.*, how the methods optimize the network (Section 5). Most deep photometric stereo networks are trained with paired photometric stereo images (input) and surface normals (ground truths), *i.e.*, supervised learning. Whether a photometric stereo network can be optimized in a self-supervised way? Whether additional information can be added to simplify the learning of surface-normal recovery? On these sides, this paper reviews recent attempts to expand and break through the supervised frameworks [30], [31], [32], [33] and gives suggestions for future developments.

Based on the classifications and summaries provided above, we then evaluate more than 30 deep learning models for photometric stereo on the widely used benchmark [6] in dense input condition (Table 3) and sparse input condition (Table 4), respectively. We found that compared with traditional non-learning methods, deep learning-based photometric stereo models are superior in estimating surface normals. Finally, we point out the future trends in the field of photometric stereo. Our aim with this survey is to help researchers understand the state-of-the-art methods and position themselves to develop in this growing field, as well as highlight opportunities in future research. The project of this survey can be found in https://github.com/Kelvin-Ju/Survey-DLCPS.

2 Problem Formulation

Consider a pixel on a non-Lambertian surface with the normal $\boldsymbol{n}$ illuminated by a directional incident light $\boldsymbol{l}$. When a linear-response camera photographs this surface in the view direction $\boldsymbol{v}$, the pixel-measured intensity m in image $\boldsymbol{M}$ can be approximated as follows:

$$m = \rho\left(\boldsymbol{n}, \boldsymbol{l}, \boldsymbol{v}\right) \cdot \max\left\{\boldsymbol{n}^{\top}\boldsymbol{l}, 0\right\} + \epsilon, \tag{2}$$

where ρ represents the BRDF, and $\max\{\boldsymbol{n}^{\top}\boldsymbol{l}, 0\}$ denotes the attached shadows, and ϵ represents global illumination effects (*e.g.*, cast shadows and inter-reflections) and noise. Traditional photometric stereo methods computed the surface normals of general objects by solving the imaging model Eq. 2 inversely, using more than three input images, but unknown BRDFs make the model difficult to fit (as shown in Fig. 1). Similarly, deep learning-based calibrated photometric stereo methods aimed to learn a neural network model f from n different observations, as follows:

$$f : \mathrm{Agg}(\boldsymbol{M^i}, \boldsymbol{l^i}) \rightarrow \boldsymbol{N}, i \in \{1, 2, \cdots, n\}, \tag{3}$$

where f is the optimized deep neural network by the training data sets. Usually, the aggregation models (Agg) are determined by how they process the input images, such as observation maps, max-pooling models, or hybrid methods, which will be discussed in Section 3. Most of the existing PS methods, *i.e.*, calibrated photometric, relied on having prior knowledge of the light directions and intensities for each image, while uncalibrated photometric stereo can estimate surface normals without lighting information. Note that the model f becomes $f : \mathrm{Agg}(\boldsymbol{M^i}) \rightarrow \boldsymbol{N}, i \in \{1, 2, \cdots, n\}$ when addressing uncalibrated photometric stereo. Although uncalibrated photometric stereo has the advantage of not requiring pre-calibration of lighting conditions, it does face additional challenges because it needs to disentangle the lighting information from shading cues, making it a more complex problem to solve. In this paper, we mainly focus on the deep learning-based calibrated photometric stereo since it provides more universal frameworks and feature extraction models that can be extended to uncalibrated and other photometric stereo tasks. A brief discussion of the uncalibrated condition can also be found in Section 3.2.4 for a more comprehensive overview.

In the following subsections, we will discuss these deep learning-based calibrated photometric stereo methods from different perspectives.

3 Categorization by Input Processing

The first deep learning method, DPSN [20], made the order of illuminations and the number of input images unchanged, by a seven-layer fully-connected network. Therefore, the following methods focused on handling any number of input images with arbitrary light directions. In fact, this problem is equivalent to how to fuse a varying number of features in the networks. It is known that convolutional neural networks (CNNs) cannot handle a varying number of inputs during training and testing. Therefore, two approaches have been proposed in photometric stereo, *i.e.*, to process the input images pixel-wise or patch-wise. Following the concept proposed in [23], we also call the pixel-wise and patch-wise processing methods as per-pixel methods (Section 3.1) and all-pixel methods (Section 3.2), respectively. We provide an in-depth summary of the development of these two approaches, in reference to the drawbacks of the initial methods (*i.e.*, the observation map from CNN-PS [21] and the max-pooling from PS-FCN [22]). In addition, we propose a new class, for hybrid methods (Section 3.3), which fuse pixel- and patch-wise characteristics. As tabulated in Table 1, we also summarize the algorithms and formulas of representative methods for each direction in Fig. 2.

3.1 Per-pixel methods

The per-pixel strategy was first implemented using the observation map in CNN-PS [21]. The observation map essentially fused all observations pixel-by-pixel, capturing the inter-image intensity variations for each pixel. Observation maps were also widely used in recent near-field photometric stereo [34] and multiview photometric stereo [35] tasks. Fig. 3 illustrates the fusion rule, which is based on both the pixel intensity and the orthogonal projected light direction. Specifically, observation maps [21] are determined by projecting light directions from a 3D space (hemisphere) onto a fixed-size observation map plane (along the axis-z direction). Each observation map can represent the feature at a single-pixel position. The observation map proves to be effective in photometric stereo for three reasons. First, its size is independent of the number of input images. Second, the values are independent of the order of the input images. Third, the information on the light directions and intensities is embedded in the observation map [21]. Recently, Ikehata [33] further took advantage of the physical interpretability of the observation map, making the observation map parse the physical intrinsic attributes to form a self-supervised inverse rendering pipeline.

3.1.1 Problem of sparse input

However, the observation map in the initial method [21] also encounters some limitations. First, light directions are represented by unstructured vectors, while observation maps are grid data as images. When projecting a light vector onto a 2D coordinate system, the projected direction can not exactly correspond to the grid observation map. To improve the accuracy of projected light directions, the size of the observation map has to be large enough to approximately represent the unstructured projected vectors. Unfortunately, the number of input images (light directions) is sparse compared to the size of the observation map, which creates difficulties in extracting features. In fact, the sparse observation map affects network performance. The accuracy of CNN-PS drops significantly when inputting a small number of images (sparse condition), compared with the all-pixel methods.

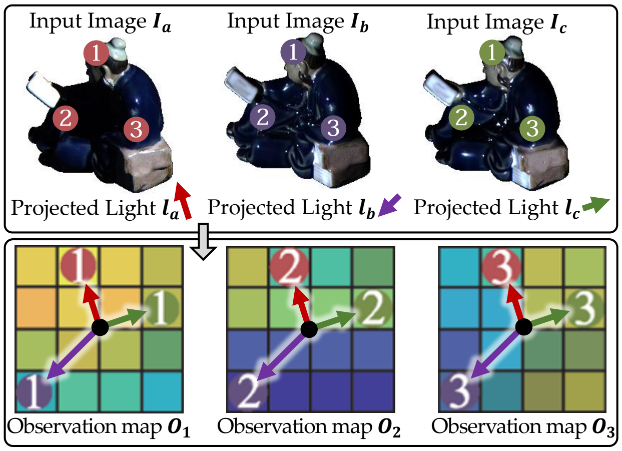


Fig. 3. The illustration of the observation maps [21]. Here, a, b, and c represent the number of input images (lights), while 1, 2, and 3 denote the index of pixel positions.

In this regard, some works were proposed to solve the sparse input images problem, such as SPLINE-Net [36] and LMPS [37]. These two methods adopted opposite strategies to solve this problem. SPLINE-Net [36] proposed a lighting interpolation network to generate dense lighting observation maps when the input was sparse (as shown in the red arrow of Fig. 4). To optimize the lighting interpolation network and normal estimation network, SPLINE-Net further utilized a symmetric loss and an asymmetric loss to consider general BRDF properties explicitly and outlier rejections, respectively. On the other hand, LMPS [37] reduced the demands on the number of images by only learning the critical illumination conditions. The method employed a connection table to select those illumination directions that were the most relevant to the surface normal prediction process (as shown in the blue arrow of Fig. 4). Furthermore, a more thorough method [38] was to replace the structured observation map with an unstructured graph network, which will be introduced in Section 3.3.

3.1.2 Problem of global information

On the other hand, since original per-pixel methods [21] operate in isolation, which means the estimated normal vector of a surface pixel relies solely on the features extracted from that pixel itself, without leveraging information from adjacent pixels. As a result, it may lose the local context information of neighboring pixels when computing the feature map.

When the input observations exhibit deviations in photometric cues, per-pixel methods might exhibit reduced robustness compared to all-pixel methods, which consider all pixels in the input patch. For example, as mentioned in [21], where the first 20 images of the “Bear” object in the DiLiGenT benchmark data set [6] were less accurate: the intensity values around the bear’s stomach region were lower than the adjacent regions, even though they should be higher due to specularities. When all 96 images of ”Bear” were fed into the per-pixel method CNN-PS [21], the mean angular error increased dramatically, from 4.20 (with the first 20 images discarded) to 8.30, an increase of 97.62%. In contrast, all-pixel methods demonstrated better robustness, *e.g.*, with PS-FCN [22], the error increased from 5.02 to 7.55, by only 50.40%. This experiment illustrates the robustness of adjacent pixels.

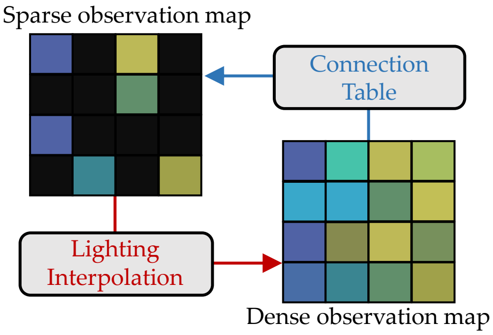


Fig. 4. Per-pixel methods for sparse input images. SPLINE-Net [36] uses the lighting interpolation network to generate dense observation maps, while LMPS [37] applies the connection table used to select the most relevant illuminant directions in the sparse observation maps.

To solve this limitation, some recent works incorporated global information into observation map-based per-pixel methods, which led to superior performance, such as PX-Net [39]. PX-Net proposed an observation map-based method that considers global illumination effects, such as self-reflections, surface discontinuity, and ambient light, which enabled global information to be embedded in the per-pixel generation process. Additionally, PX-Net performed well in handling sparse conditions, in contrast to the original observation map-based method [21]. Other methods, such as HT21 [40] and GPS-Net [38], learned global information (intra-image features) by combining the per-pixel and all-pixel strategies. We will discuss these methods in Section 3.3.

3.2 All-pixel methods

In contrast to per-pixel methods, which analyze each pixel individually in observations, all-pixel methods keep all the pixels together. All-pixel methods have the advantage of exploring intra-image intensity variations across an entire input image. The original all-pixel method was introduced in PS-FCN [22] through the use of a max-pooling layer, which operated in the channel dimension and fused features from an arbitrary number of inputs. At each position in the fused feature, the value was determined as the maximum among all the input features at that position. Consequently, this method allowed a convolutional network to work with features from any number of inputs. The max-pooling layer was inspired by aggregating multi-image information in other computer vision tasks [41], [42]. Compared to variable input methods like RNN [43], the adopted max-pooling operation was order-agnostic, meaning it was not sensitive to the order in which the input images were provided. This attribute made it particularly suitable for photometric stereo. The all-pixel max-pooling operation offers several advantages. Firstly, it can handle an arbitrary number of input images without being affected by their order. Secondly, the use of whole image features includes valuable local context information, which enhances surface normals estimation. Thirdly, the patch-based input accelerates the training process compared to per-pixel methods. Lastly, all-pixel methods handle the input images and lighting

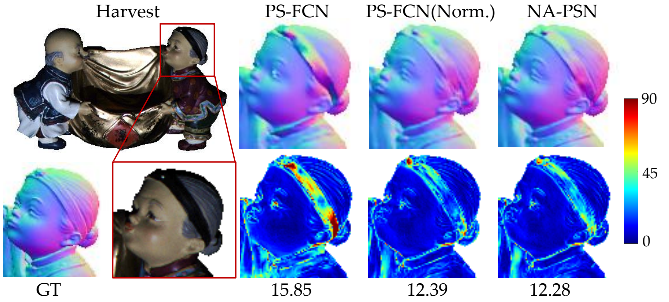


Fig. 5. Examples of the predictions and error maps on spatially varying BRDF, from the object “Harvest” in the DiLiGenT data set [6]. NA-PSN is short for NormAttention-PSN. The numbers reveal the mean angular error in degrees.

directions (as extra information) separately, making them capable of predicting photometric stereo under unknown illuminations (uncalibrated photometric stereo).

3.2.1 Problem of spatially varying BRDF

However, the original all-pixel method PS-FCN [22] had some limitations. To begin with, PS-FCN cannot handle surfaces with spatially varying materials. Since all-pixel methods leverage convolutional networks to process input in a patch-based manner, they may have difficulties in dealing with steep color changes caused by surfaces with spatially varying materials. It can be seen as the negative effect of considering observations in the neighborhood when computing the feature maps. As shown in Fig. 5, the head and collar region is with spatially varying BRDF. The original per-pixel method PS-FCN [22] was less effective in handling regions with spatially varying BRDF, where the color change of the beard influenced the surface normal map. While improved methods, such as PS-FCN (Norm.) [44] and NormAttention-PSN [45], showed significantly enhanced reconstruction results. This problem may be rooted in two key factors. Firstly, the feature extraction network encounters difficulty in decoupling the changes between the photometric shading cues and BRDFs. In other words, the feature extraction network may struggle to differentiate between changes in pixel values due to variations in surface structures and those resulting from different material properties. Secondly, the per-pixel methods inherently incorporate local context information, where each estimated surface normal vector depends on neighboring pixels when computing feature maps. Consequently, surface normal estimations can be influenced by spatially varying BRDF.

To solve this limitation, Chen *et al.* further proposed PS-FCN (Norm.) [44]. Rather than creating a large-scale training set with spatially varying materials, an observation normalization method, which concatenated all the observations and normalizes them, was introduced, as follows:

$$m_i' = \frac{m_i}{\sqrt{m_1^2 + \cdots + m_n^2}},\ i \in \{1, 2, \cdots, n\}, \tag{4}$$

where m_i and m_i' represent the original and normalized pixel intensities in the n images. Under the assumption of Lambertian reflectance, the effect of albedo can be removed. However, PS-FCN (Norm.) [44] cannot perfectly handle the condition of non-Lambertian surfaces. In regions with specular highlights, the denominator of Eq. 4 becomes larger, leading to the suppression of observations after normalization [44]. As shown in the red boxes in Fig. 6, the original normalization method [44] excessively suppressed the highlighted regions, whereas the double-gate normalization method [45] provided more reasonable shading cues in these regions. Although max-pooling can naturally ignore non-activated features, the suppressed observations are not equal to the suppressed features, *i.e.*, the changing appearance of an observation may cause larger feature values. Therefore, Ju *et al.* [45] proposed a double-gate observation normalization to better handle the non-Lambertian surfaces with spatially varying materials. In the method, two gates were set at the lowest 10% (P_{10}) and the highest 10% (P_{90}) grayscale values of all pixels and put them on the denominator of Eq. 4, as follows:

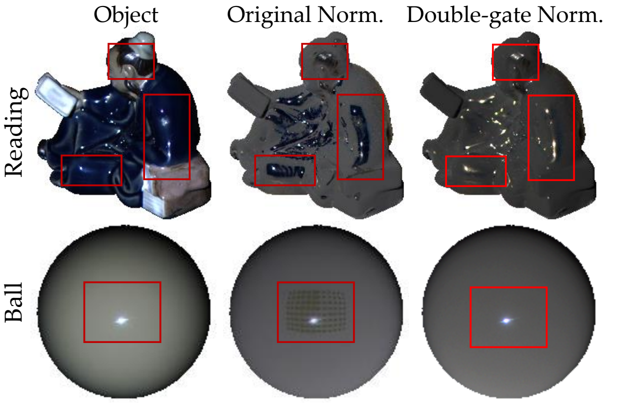


Fig. 6. Comparison of the original normalization method [44] and the double-gate normalization method [45], with the input object “Reading” and “Ball”. The red boxes represent the regions exhibiting specular highlights.

$$m_i' = \frac{m_i}{\sqrt{\sum_k m_k^2}},\ k \in \mathcal{S}, \tag{5}$$

where the set $\mathcal{S}$ is controlled by the two gates, such that $m_i \in \mathcal{S}$ if $Gate(P_{10}) < m_i < Gate(P_{90})$, for $i = 1, 2, \cdots, n$. It can be seen that the non-Lambertian effects are removed in the red boxes in Fig. 6. However, this method has to concatenate with the original images, since discarding some grayscale values in the denominator can be viewed as a nonlinear process, which may affect the shading cues for photometric stereo [45].

3.2.2 Problem of blurry details

The second limitation of original all-pixel methods is that they may cause blurred reconstructions in complex-structured regions. We believe that the reasons mainly lie in three. (1) The convolutional models process patch-based input, which means that all normal points will mutually affect each other and cause blurring, especially in high-frequency areas. (2) The widely used Euclidean-based loss functions can hardly constrain the high-frequency (*i.e.*, complex-structured) representations, because of the ”regression-to-the-mean” problem [46], which results in blurry and over-smoothed images. (3) Previous network architectures pass the input through high-low-high resolutions, *i.e.*, through an encoder-decoder architecture, which leads to the loss of prediction details and causes blurring.

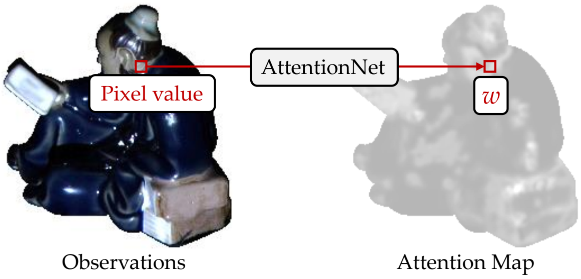


Fig. 7. An example of an attention map from Attention-PSN [47]. w is the weight of the pixel-wise attention-weighted loss (Eq. 6).

In this regard, two different strategies were proposed to deal with the problem of blurred reconstruction in all-pixel methods. The first approach was to employ adaptive loss for different kinds of surfaces. Attention-PSN [47] was the first to propose an attention-weighted loss to produce detailed reconstructions, as follows:

$$\mathcal{L} = w\mathcal{L}_{\text{gradient}} + (1 - w)\mathcal{L}_{\text{normal}}, \tag{6}$$

which learned a higher weight (w) for the detail-preserving gradient loss $\mathcal{L}_{\text{gradient}}$ and a lower weight $(1 - w)$ for the cosine loss $\mathcal{L}_{\text{normal}}$ for high-frequency regions. As shown in Fig. 7, Attention-PSN [47] learned an attention map from input images, whose pixel values became the weights of the attention-weighted loss. However, the surface materials of an object may change rapidly in a flat or smooth region, which affects the gradient loss with a large weight in the region and dilutes the penalty on surface normals. Therefore, Ju *et al.* further employed the above double-gate observation normalization to eliminate the influence of spatially varying surface materials, namely NormAttention-PSN [45].

On the other hand, the second approach was to preserve the high-resolution features via novel network architectures. CHR-PSN [48] proposed a parallel network structure for maintaining both deep features and high-resolution details of surface normals, inspired by the High-resolution Net (HR-Net) [26] for human pose estimation. Full-resolution features can always be preserved in the network, avoiding features passing layers from high to low resolution and blurring.

3.2.3 Problem of fusion efficiency

The third is that the fusion mechanism of all-pixel methods, *i.e.*, max-pooling, discards a large number of features from the input, reducing the utilization of information and affecting the estimation accuracy. Therefore, how to retain more features with key information is essential. Some methods [38], [49] fused max-pooling and average pooling via a concatenation operation. However, the improvement of adding average-pooling is limited because averaging features may smooth out saliency and dilute valuable features. Different from adding averaging information, Manifold-PSN [50] introduced nonlinear dimensionality reduction [51] to convert features from high-dimensional feature spaces to low-dimensional manifolds. However, the manifold method truncated the backpropagation of the network. Therefore, the authors had to use the max-pooling layer to pre-train the extractor of the network, which was cumbersome and inefficient.

On the other hand, some methods employed novel models to enhance feature fusion in their structures. MF-PSN [52] introduced a multi-feature fusion network, utilizing max-pooling operations at different feature levels in both shallow and deep layers to capture richer information. Besides, CHR-PSN [48], SR-PSN [53], and MS-PS [54] extended max-pooling at various scales with different receptive fields, rather than the depth. Furthermore, HPS-Net [55] introduced a bilateral extraction module that generated positive and negative information before aggregation to better preserve useful data. Despite these advancements in feature fusion, none of these methods fully address the essential challenge of information loss, *i.e.*, the max-pooling layer only extracts the maximum value, ignoring the rest.

Recently, the Transformer architecture [27] has also been used to fuse and communicate features from different input images. PS-Transformer [56] first used a multi-head attention pooling [57] to fuse an arbitrary number of input features. In this way, the number of elements in a set was shrunk from an arbitrary dimension to one, by giving a learnable query Q rather than only retaining the maximum value. Multi-head attention pooling [57] can be seen as a global fusion method that considers all feature distributions, instead of only retaining the maximum value.

3.2.4 Uncalibrated condition

Most of the existing methods, *i.e.*, calibrated photometric stereo, require knowledge of the light direction and intensity for each image. However, calibrating the light involves complex operations and relies on specialized instruments, which may make it impractical for real-world applications. In contrast, uncalibrated photometric stereo can estimate surface normals without requiring lighting information. However, it encounters more challenges, such as the Generalized Bas-Relief (GBR) ambiguity [58] and general non-Lambertian surface reflectance.

As discussed in Section 3.1, per-pixel methods rely on the projected light direction from 3D space onto the 2D observation map, where light directions are essential. Conversely, all-pixel methods handle input images and light directions separately. Therefore, the all-pixel strategy is first naturally applied in uncalibrated conditions. The per-pixel method PS-FCN [22] first addressed the uncalibrated problem by directly learning the mapping from input images to surface normals without concatenating light directions, denoted as UPS-FCN. However, the performance of UPS-FCN is far from satisfactory due to the complex interplay among shading cues, which include unknown lighting directions, surface normals, and reflectance properties. To address the uncalibrated case more effectively, most deep learning-based uncalibrated photometric stereo methods adopted a two-stage strategy. This involves first estimating the light directions and then estimating surface normals using both the estimated light information and input images, based on all-pixel networks [59], [60], [61], [62], [63], [64].

SDPS-Net [59] first proposed the two-stage deep learning architecture to reduce the learning difficulty in uncalibrated photometric stereo. It began by estimating light directions and intensities via the light calibration network, then applied an all-pixel-based normal estimation network to obtain the surface normal map. UPS-GCNet [60] used object

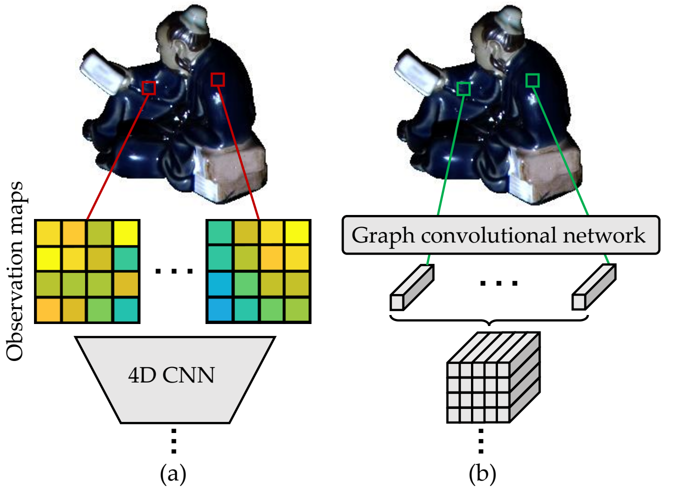


Fig. 8. Hybrid methods for handling both per-pixel features and all-pixel features. (a) Schematic of the 4D CNN to capture the global effect of the observation maps used in HT21 [40]. (b) Schematic of the structure-aware Graph Convolution network to extract a fixed-size feature map, used in GPS-Net [38].

shape and shading information as guidance to improve lighting estimation. Similarly, ReDDLE-Net [62] incorporated diffuse and specular cues to enhance light estimation. Sarno *et al.* [61] employed differentiable neural architecture search (NAS) to automatically discover the most efficient neural architecture for both light calibration and normal estimation networks. In addition to supervised methods, a few uncalibrated methods were implemented in self-supervised and multi-supervised ways. For example, Kaya *et al.* [65] used an uncalibrated neural inverse rendering approach to handle unknown lights, and Li *et al.* [64] allowed re-rendered errors to be back-propagated to the light sources and refined them jointly with the normals. Yang *et al.* [66] utilized the neural reflectance field to realize the 3D reconstruction from uncalibrated photometric stereo images with the capability of recovering invisible parts. Tiwari *et al.* [63] jointly trained the network with image relighting and used multiple loss functions to optimize the network.

3.3 Hybrid methods

As discussed above, both per-pixel and all-pixel methods come with their own sets of advantages and limitations. Per-pixel methods primarily focus on analyzing inter-image intensity variations at the pixel level. In contrast, all-pixel methods pay more attention to extracting features related to intra-image lighting variations. Hybrid approaches that combine these strategies may have the benefits of both per-pixel and all-pixel techniques.

In fact, the first mixed method can be found in learning-based multispectral photometric stereo [67], which initially estimated a coarse surface normal map and subsequently refined it using a per-pixel approach, achieved through a fully connected network. Recently, MT-PS-CNN [68] proposed a two-stage photometric stereo model to construct inter-frame (per-pixel) and intra-frame (all-pixel) representations. Similarly, Yang *et al.* [69] introduced a tandem manner for per-pixel and all-pixel feature extraction, namely PSMF-PSN. This network employed 3D convolutional layers to extract pixel-wise features. In addition, PS-Transformer [56] introduced a dual-branch feature extractor based on the self-attention mechanism [27], exploring both pixel- and image-wise features. Honzatko *et al.* [40] built upon the observation maps but incorporated spatial information using 2D and 4D separable convolutions to better capture global effects. Differently, GPS-Net [38] introduced a structure-aware graph convolutional network [70] to establish connections between an arbitrary number of observations per pixel, without relying on observation maps. Subsequently, convolutional layers were employed to extract spatial information. These hybrid methods may benefit from per-pixel and all-pixel approaches. As shown in Fig. 8, we summarize the hybrid strategy of HT21 [40] and GPS-Net [38].

However, all existing hybrid methods follow a sequential and independent approach to extracting per-pixel and all-pixel features. Future research may focus on effective ways of combining these two feature types and consider the learning process as a holistic approach, rather than treating it as two separate stages.

4 Network Architectures

With the development of deep learning techniques, deep learning-based photometric stereo networks have used many advanced modules. In this Section, we will review these modules and compare their advantages and drawbacks in the task of surface normal recovery.

4.1 Convolutional networks

In the beginning, DPSN [7], [20] utilized Multilayer Perceptron (*i.e.*, fully connected layers) with dropout layers to map the surface normals from observations pixel by pixel. However, this architecture ignores adjacent information and cannot handle a flexible number of input images. Therefore, PS-FCN [22] and CNN-PS [21] were proposed to handle an arbitrary number of input images by different strategies (max-pooling and observation map). PS-FCN [22] applied a fully convolutional plain network to learn surface normals, while CNN-PS [21] used a variant of the DenseNet architecture [25] to estimate surface normals from an observation map. The DenseNet architecture [25] has been widely used in subsequent networks, such as LMPS [37], SPLINE-Net [36], MF-PSN [52], and PX-Net [39], due to its excellent feature extraction capacity. Similarly, ResNet [24] was also widely used in deep learning-based photometric stereo methods [38], [47], [50], [71], which can effectively avoid gradient vanishing in deep networks. However, the above structures ignore keeping the high resolution of the features, *i.e.*, passing the features sequentially from high-to-low resolution layers, and then increasing the resolution. This operation is suitable for a high-level task that needs semantic features. However, it may cause information loss and blurring for the per-pixel prediction photometric stereo task. Therefore, some works [45], [48] introduced a parallel multi-scale structure, inspired by the improvement of HR-Net [26] in the human pose estimation task. HR-Net [26] employed a parallel network structure to extract features at three scales, avoiding the feature map being changed from low resolution to high resolution, where the feature

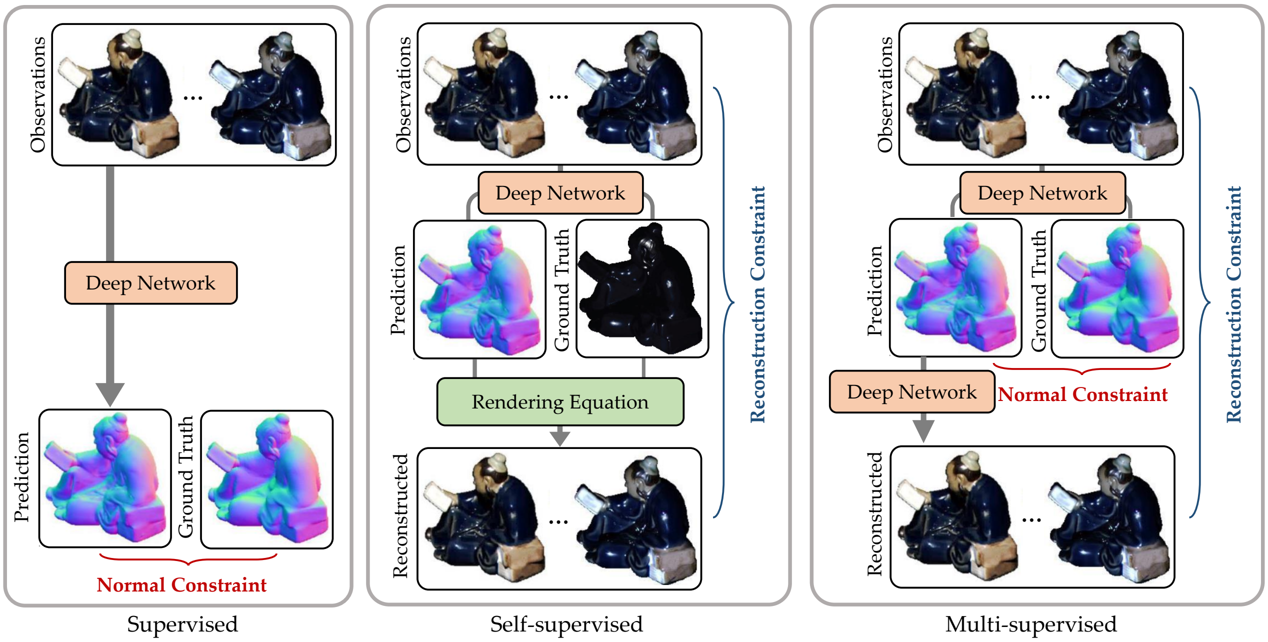


Fig. 9. Comparison of the frameworks of supervised, self-supervised, and multi-supervised photometric stereo.

extraction process maintained both the deep features with high semantic and high-resolution features having details for surface-normal prediction.

4.2 Self-attention mechanism

Transformer with a self-attention module [27] was first proposed in the field of natural language processing. It has also been widely used in many computer vision tasks, where self-attention was employed in the spatial dimensions to capture non-local feature dependencies. Recently, two works [56], [72] introduced the self-attention mechanism to aggregate features under different lights in the context of photometric stereo. SPS-Net [72] was the first to propose a self-attention photometric stereo network, which aggregated photometric information through a self-attention mechanism. Ikehata *et al.* [56] then presented PS-Transformer, which uses the self-attention mechanism to capture complex interactions in sparse photometric stereo. PS-Transformer [56] designed a dual branch to explore pixel and image-wise features. Therefore, intra-image spatial features and inter-image photometric features are better extracted than with SPS-Net [72]. Recently, Ikehata introduced two photometric stereo methods: UniPS [73] and SDM-UniPS [74]. These methods can handle natural lighting conditions by learning global lighting contexts from individual images by interacting with others, discarding assuming specific lighting models. In these approaches, the self-attention model [27] served as the backbone to facilitate non-local interactions and as the aggregation method to fuse arbitrary features.

4.2.1 Discussion

The Transformer module showed significant performance improvements in other computer vision domains [75], [76]. Similarly, the photometric stereo task can leverage the self-attention module effectively. Theoretically, the surface normal of a point only depends on itself, rather than its relationship with distant points. However, due to the presence of shadows and inter-reflections, capturing long-range context becomes essential for accurate feature extraction. Therefore, Transformer-based photometric stereo models can benefit from both the non-local information acquired through the self-attention module and the embedded local context information obtained through traditional convolutional layers. Furthermore, the effectiveness of the Transformer [56], [73], [74] can facilitate communication and aggregation of features flexibly, by using multi-head attention pooling [57].

However, Transformer-based photometric stereo methods face some limitations. The Transformer module has greater modeling flexibility and can focus on the information at any position. Consequently, in contrast to convolutional networks, it requires larger-scale training data sets. Furthermore, it is widely recognized that the Transformer model imposes substantial computational demands, especially when dealing with a large number of elements [56]. Hence, future research should explore the adaptation of Transformer-based photometric stereo methods to address dense problems more efficiently.

5 Categorization by Supervision

As a mapping task, conventional learning-based photometric stereo methods optimized the network by minimizing the distance between predicted surface normals and ground-truth surface normals [20], [21], [22], supervised by pairs of photometric stereo images and their surface normals. However, learning-based 3D tasks face challenges due to the difficulties in acquiring and aligning a large number of ground truths. To solve this issue, some researchers have investigated self-supervised learning in photometric stereo [30], [65], [77]. Besides, many works further improved the performance by introducing additional supervision [32], [33], [78] or additional information to simplify optimization [31], [71], [79]. In Fig. 9, we summarize the differences among supervised, self-supervised, and multi-supervised photometric stereo networks.

5.1 Supervised photometric stereo methods

Plenty of deep photometric stereo networks have been proposed with improved performance, compared to traditional handcrafted photometric stereo methods. These learning-based models show the potential ability of deep neural networks with supervised optimization, *i.e*, a large amount of data with ground-truth surface normals during the training stage. Among these supervised models, some methods [7], [20], [21], [37] utilized the L2 loss (*i.e.*, mean squared error loss), as follows:

$$\mathcal{L} = \|\boldsymbol{n_p} - \tilde{\boldsymbol{n}}_{\boldsymbol{p}}\|_2^2, \tag{7}$$

while more methods applied cosine similarity loss, as follows:

$$\mathcal{L} = 1 - \boldsymbol{n_p} \odot \tilde{\boldsymbol{n}}_{\boldsymbol{p}}, \tag{8}$$

where $\odot$ represents the dot-product operation. In this case, $\boldsymbol{n_p} \odot \tilde{\boldsymbol{n}}_{\boldsymbol{p}}$ will be close to 1 when the predicted $\tilde{\boldsymbol{n}}_{\boldsymbol{p}}$ is similar to the ground truth $\boldsymbol{n_p}$, and Eq. (8) will approach 0. Intuitively, the cosine similarity loss is more suitable for surface-normal estimation, because it directly measures the difference in orientation between two vectors. However, no evidence from previous work shows that L2 loss reduces the accuracy of estimated surface normals with the same network architecture and settings.

5.1.1 Additional information

Recently, some supervised photometric stereo networks improved performance with additional information to make optimization more efficient [31], [71], [79]. The additional information can be regarded as prior knowledge used to simplify the optimization of deep networks through weight parametrization. In contrast to previous deep learning approaches that solely derived the normal space from the observed shading cues, these methods leveraged both additional information and observations to learn the surface normal. Consequently, these methods had the capacity to reduce the learning hypothesis space, leading to easier feature extraction, faster convergence, and improved learning accuracy.

Wang *et al.* [31] proposed a non-Lambertian photometric stereo network with the additional collocated light image. Their model leveraged the monotonicity of isotropic reflectance and the univariate property of the supplementary collocated light to facilitate the decoupling of the surface normal from the reflectance function, in conjunction with the input photometric stereo images. Ju *et al.* [71] incorporated initial normal priors to enhance the accuracy of surface normal predictions for objects. This approach relied on prior surface normals based on Lambertian assumption [1] to reparameterize network weights, enabling the alignment of mappings in the same normal space and increasing the focus on the errors in the prior normal. Similarly, Ju *et al.* [79] proposed an additional reflectance-guided photometric stereo network, which employed a dual-branch extractor to combine information from both prior reflectance and photometric stereo images. Furthermore, the inclusion of prior reflectance helped eliminate the impacts of surfaces with spatially varying reflectance for photometric stereo methods. These methods can enhance performance by incorporating additional information to streamline the optimization process.

In general, supervised photometric stereo methods can achieve superior performance, but these methods are limited due to the difficulties in acquiring accurate ground truth for the photorealistic training sets, and there is a gap between real photo images and synthetic images due to rendering techniques.

5.2 Self-supervised photometric stereo methods

As discussed above, deep learning techniques drastically advanced the photometric stereo task. Current existing deep learning methods usually solve the problem in a supervised training manner. These methods relied on a large amount of training data with ground truth. However, measuring the surface normals of real objects is very difficult and expensive, because it needs high-precision 3D scanners to reconstruct the ground-truth shape, and requires much manpower to align the viewpoints between surface normal maps and multiple images (pixel alignment). Until now, only three real-world scene data sets with ground truth have been proposed [6], [80], [81]. However, these data sets only contained 10 to 100 object scenes and were far from being utilized for training a modern deep neural network. Synthetic training data is a possible way [20], [21], [22], but synthetic images should account for various realistic BRDF, object shapes, cast shadows, and inter-reflections, *etc*. Existing BRDF databases [28], [29] and renderers still required efforts to generate photo-realistic synthetic images.

To overcome the above shortcomings, some researchers introduced the self-supervised learning strategy, which only needs photometric stereo images, rather than pairs with ground truth surface normals [30], [65], [77]. The pipeline of the self-supervised photometric stereo methods can be described as Eq. 9, as follows:

$$\boldsymbol{M}_R = \Psi(\Phi(\boldsymbol{M})), \tag{9}$$

where Φ represents the model responsible for extracting the surface normal, while Ψ denotes the method used to re-render the reconstructed images $\boldsymbol{M}_R$. In this case, the estimated surface normal $\tilde{\boldsymbol{N}} = \Phi(\boldsymbol{M})$ is optimized by minimizing the reconstruction loss (*e.g.*, L2 loss) between the input images $\boldsymbol{M}$ and the re-rendered images $\boldsymbol{M}_R$, without requiring the ground-truth surface normal of $\boldsymbol{M}$.

Taniai and Maehara [30] first proposed a self-supervised convolutional network that took the whole set of images as input, namely IRPS. The model directly generated surface normals by minimizing the reconstruction loss between re-rendered images obtained via the rendering equation and input images. Furthermore, IRPS [30] avoided the unfixed number of input photometric images through its physics-based rendering approach. However, IRPS suffered from expensive computation [23] and failed to model inter-reflections. Moreover, its loss function was not robust and susceptible to noise [65] because the surface normals were initialized by using the Lambertian assumption [1]. Therefore, IRPS was further extended by Kaya *et al.* [65] to deal with inter-reflection by explicitly modeling the concave and convex parts of a complex surface. However, both [30], [65] implicitly encoded the specular components as features of the network and fail to consider shadows in the rendering process. To solve the limitations, Li *et al.* [77]

TABLE 1
Representative calibrated deep learning photometric stereo algorithms and formulations for each taxonomy.

Taxonomy	Sec.	Method	Formulation
Fixed input	3	DPSN [20]	$\boldsymbol{n_p} = f(\tilde{m}_p^i)$, where $\tilde{m}_p^i$ means the order and the number of the inputs are fixed.
Per-pixel input	3.1	CNN-PS [21]	$\boldsymbol{n_p} = f(\mathrm{obs}(m_p^i, l^i))$, please see obs in Section 3.1.
Sparse input	3.1.1	LMPS [37]	$\boldsymbol{n_p} = f(D(\mathrm{obs}(m_p^i, l^i)))$, where D stands for the connection table.
Global information	3.1.2	PX-Net [39]	$\boldsymbol{n_p} = f(\mathrm{obs}(G(m_p^i), l^i))$, where G stands for the global illumination effects.
All-pixe input	3.2	PS-FCN [22]	$\boldsymbol{N} = f_R(\max\{f_E(\boldsymbol{M^i}, l^i)\})$, please see max in Section 3.2, f_E and f_R mean the Extractor and Regressor.
Spatially varying BRDFs	3.2.1	PS-FCN (Norm.) [44]	$\boldsymbol{N} = f_R(\max\{f_E(\mathrm{norm}(\sum_i^n(\boldsymbol{M^i})), l^i)\})$, where norm stands for the Observation Normalization.
Blurry details	3.2.2	Attention-PSN [47]	$\boldsymbol{N} = f_R(\max\{f_E(\boldsymbol{M^i}, l^i)\})$, with minimizing the adaptive loss $\mathcal{L} = \lambda\mathcal{L}_{\mathrm{gradient}} + (1-\lambda)\mathcal{L}_{\mathrm{normal}}$.
Fusion efficiency	3.2.3	MF-PSN [52]	$\boldsymbol{N} = f_R(\max_d\{\max_s\{f_E(\boldsymbol{M^i}, l^i)\}, f_E(\boldsymbol{M^i}, l^i)\})$, where s and d mean shallow and deep, respectively.
Hybrid input	3.3	GPS-Net [38]	$\boldsymbol{N} = f(\cup_p^{H\times W}\mathrm{gcn}(m_p^i, l^i))$, where gcn stands for the Structure-aware Graph Convolution filters.
Self-attention	4.2	PS-Transformer [56]	$\boldsymbol{N} = f_R(f_E(m_p^i, l^i), f_E(\boldsymbol{M^i}))$, where f_E and f_R are layers with the Self-attention Mechanism.
Additional information	5.1.1	WZ20 [31]	$\boldsymbol{n_p} = f_R(\max\{f_E(m_p^i, l^i, m_p^0\})$, where m_p^0 stands for the collocated light observation.
Self-supervised	5.2	IRPS [30]	$\boldsymbol{N} = f_P(\boldsymbol{M^i}) = f_I^{-1}(\boldsymbol{M^i})$, with minimizing the self-supervised loss $\mathcal{L} = f_I(f_P(\boldsymbol{M^i}), l^i) - \boldsymbol{M^i}$.
Multi-supervised	5.3	DR-PSN [78]	$\boldsymbol{N} = f_P(\boldsymbol{M^i}, l^i) = f_I^{-1}(\boldsymbol{M^i})$, where f_P and f_I mean the Normal regression and Dual regression.

proposed a coordinate-based deep network to parameterize the unknown surface normal and the unknown reflectance at every surface point. The method learned a series of neural specular basis functions to fit the observed specularities and explicitly parameterized shadowed regions by tracing the estimated depth map. However, the method may fail in the presence of strong inter-reflections.

In summary, self-supervised photometric stereo models alleviate the demand for extensive 3D data sets. However, these self-supervised models face computational burdens due to their relatively large parameter sizes, which may restrict their applicability in industrial settings. We envision that future progress in self-supervised photometric stereo will involve the refinement of rendering equations for increased accuracy, the creation of more lightweight models, and the enhancement of the efficiency of reconstruction loss.

5.3 Multi-supervised photometric stereo methods

Previous research [31], [71] demonstrated improved performance by incorporating additional input information within supervised learning frameworks. Another approach to simplifying the learning process is to introduce more forms of supervision. In this paper, we refer to methods that utilize multiple forms of supervision as "Multi-supervised" [32], [33], [78].

In this context, Ikehata [33] proposed a network to deconstruct the observation map into physical interpretable components such as surface normal, surface roughness, and surface base color. These components were then integrated via the physical formation model [82]. Consequently, the training loss for optimization consisted of the normal reconstruction loss and the inverse rendering loss. On the other hand, Ju *et al.* [78] introduced a dual regression network for calibrated photometric stereo, known as DR-PSN. This network combined the surface-normal constraint with the constraint of the reconstructed re-lit image. Additionally, GR-PSN [32] utilized a parallel framework to simultaneously learn two arbitrary materials for an object and included an additional material transform loss. These methods employed an inverse subnetwork to re-render reconstructed images based on predicted surface normals. In contrast to previous inverse rendering methods [30], [33], [65], [77], DR-PSN and GR-PSN used CNNs to render reconstructed images rather than following the rendering equation.

Finally, based on the taxonomy discussed in Sections 3, 4, and 5, we formulate these representative deep learning-based calibrated photometric stereo methods in Table 1. In these formulas, the predicted surface normals are represented by $\boldsymbol{N}$ from input photometric images $\boldsymbol{M^i}$ or $\boldsymbol{n_p}$ from pixel m_p^i, where p stands for the index of spatial resolution $H \times W$, $i \in \{1, 2, \cdots, n\}$ stands for the index of inputs. f represents deep neural networks for learning surface normals.

6 Data Sets of Photometric Stereo

The training and testing of supervised photometric stereo networks require the ground truth normal maps of objects. However, obtaining ground-truth normal maps of real objects is a difficult and time-consuming task. Although many data sets have been established in other 3D reconstruction tasks [85], [86], [87], most of their objects were simple in reflectance and shape, and the number of different lighting conditions was small [6]. This section will review data sets for deep learning-based photometric stereo methods and summarize them in Table 2. It is worth noting that we mainly review these data sets for directional lighting photometric stereo methods. Those for multiview photometric stereo [88], near-field photometric stereo [89], and multispectral photometric stereo [90] are not discussed here.

6.1 Training data sets

Training a deep photometric stereo network needs to render plenty of materials, geometries, and illumination. Researchers have to establish synthetic training data sets by rendering 3D shapes with different reflectance. There are two mainstream data sets.

6.1.1 *Blobby and Sculpture data set*

The Blobby and Sculpture data set was proposed in DPSN [20] and improved in PS-FCN [22]. PS-FCN applied 3D shape models from the Blobby shape data set [91] and the Sculpture shape data set [41], as well as the MERL BRDF data set [28] to provide surface reflectance. The Blobby shape data set contained ten objects with various shapes. The Sculpture shape data set [41] further provided more complex (detailed) normal distributions for rendering. The MERL BRDF data set [28] contained 100 different BRDFs

TABLE 2
Summary of data sets for deep learning photometric stereo.

Data set	BRDF	Ground Truth	Number of Sample	Trained methods
Blobby and Sculpture [22]	MERL [28], homogeneous	Synthetic normal	85212	[22], [37], [44], [47], [50], [68], [71] [31], [38], [45], [48], [52], [72], [78]
CyclePS [21]	Disney [29], spatially varying		45 (75 in [56])	[21], [36], [39], [40], [56]
Gourd & Apple [83]	Real object, spatially varying	Not provide	3	Test sets
Light Stage Data Gallery [84]			9	
DiLiGenT [6]		3D scanner	10	
DiLiGenT-II [81]			30	
DiLiGenT10^2 [80]	Real object, homogeneous	CAD + CNC	100	

with real-world materials, which can provide a diverse set of surface materials for rendering shapes. The authors used a physically based ray tracer, Mitsuba [92] to render photometric stereo images. For each selected shape in these two shape data sets [41], [91], the authors of PS-FCN [22] used 1296 regularly-sampled views, randomly selected 2 of the 100 BRDFs in the MERL BRDF data set, and 64 light directions randomly sampled from the upper hemisphere space to render 64 photometric stereo images with a (cropped) spatial resolution of 128 × 128. The total number of training samples for the first method was 85212.

6.1.2 CyclesPS data set

The second one was proposed in CNN-PS [21], namely the CyclesPS data set. In this data set, the authors utilized Disney's principled BSDF data set [29] rather than the MERL BRDF data set [28] to provide surface reflectance. Compared to the MERL BRDF data set [28], which had 100 measured BRDFs and thus cannot cover the tremendous real-world materials, Disney's principled BSDF data set integrated five different BRDFs controlled by 11 parameters, which can represent a wide variety of real-world materials. Although the CyclesPS data set neglected some combinations of parameters that were unrealistic or did not strongly affect the rendering results, Disney's principled BSDF data set can represent almost infinite surface reflectance. The number of 3D model shapes was 15, selected from the Internet under a royalty-free license. The CyclesPS data set [21] used the Cycles renderer, bundled in Blender [93], to simulate complex light transport, and included three subsets, diffuse, specular, and metallic. Therefore, the total number of samples for training was 45. Different from PS-FCN [22], it divided the object region in the rendered image into 5000 superpixels and used the same set of parameters at the pixels within a superpixel, *i.e.*, 5000 kinds of materials in one sample. Moreover, the number of light directions was 740, which means that 740 photometric stereo images were rendered for each sample, with a spatial resolution of 256 × 256. The number of objects in the CyclesPS data set was further increased to 25 in PS-Transformer [56], with the same settings.

6.1.3 Discussion

Compared to these two data sets, we can find that the strategy and attention are quite different. As summarized in Table 2, the Blobby and Sculpture data set [22] contains much more samples than the CyclePS data set [21] (85212 vs. 45). However, the number of illuminated images with homogeneous reflectance is 64 in the Blobby and Sculpture data set [22], while there are more than 700 very densely illuminated images with spatially varying materials in the CyclePS data set [21]. The Blobby and Sculpture data set [22] is more suitable for all-pixel methods (see Section 3.2), and the CyclePS data set [21] is better to be used by per-pixel methods (see Section 3.1). There are two reasons. First, all-pixel methods process input images in a patch-wise manner. In contrast, per-pixel methods use the observation map to learn the feature of a single pixel. Therefore, the number of samples is irrelevant as long as the spatial resolution is large enough. Second, before the introduction of the observation strategy [44], all-pixel methods with patch-based inputs cannot handle objects with spatially varying materials, while per-pixel methods naturally avoid this problem. Therefore, previous per-pixel methods usually chose the CyclePS data set [21] to optimize their models, while all-pixel methods always used the Blobby and Sculpture data set [22] (Tabulated in Table 2). However, the diversity of CyclePS [21] is much better due to the powerful representation ability of Disney's principled BSDF data set [29], which can potentially lead to better performance of per-pixel methods using the observation maps strategy. Thus, establishing Disney's principled BSDF [29] based data sets with more samples is important and urgent in future work.

6.1.4 Settings and implementation details

When training photometric stereo networks, the preferred optimizer was typically the Adam optimizer [94] with default settings ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) due to its excellent performance and ease of parameter tuning. However, some networks, such as HT21 [40], opted for the RMSprop optimizer [95].

For per-pixel methods, the size of observation maps during training can influence optimization performance. In CNN-PS [21], the authors tested and found that using observation maps with the size of 32 × 32 resulted in the best performance. Consequently, subsequent per-pixel methods usually adopted this size, or a slightly larger size (*e.g.*, 48 × 48 in HT21 [40]) in their training configurations. Similarly, for all-pixel methods, the size of input patches during training impacts performance and training time. Usually, the default size for input patches in all-pixel methods was set to 32 × 32 to achieve optimal results. MF-PSN [52] made a quantitative comparison of performance using different input sizes, supporting the choice of a 32 × 32 patch size as it can best balance both performance and computational costs.

Photometric stereo networks are required to handle a varying number of inputs. Consequently, choosing the number of inputs during training is also an important setting that has to be discussed. The experiments in [45], [52] have demonstrated that different number of inputs during training impacts the performance of photometric stereo networks. Specifically, when the number of input images used for training is close to the number used for testing, the networks will achieve better performance. In order to accommodate both sparse and dense input conditions, all-pixel methods commonly select a training input number of 32 images. In contrast, per-pixel methods often use a larger number of inputs, such as 50 to 1300, in many models that rely on observation maps. However, there are exceptions for methods specially designed for sparse inputs, such as SPLINE-Net [36], which utilizes 10 inputs.

6.2 Testing data sets

Test data sets are also needed to quantitatively evaluate the performance of different photometric stereo methods. These data sets can be divided into two categories: synthetic data sets and real data sets.

6.2.1 Synthetic data sets

Synthetic data sets were usually rendered with the same settings as the Blobby and Sculpture data set [22] or the CyclePS data set [21]. For example, the rendered objects “Bunny”, “Dragon”, and “Armadillo” in the Stanford 3D data set [96] by the MERL BRDF data set [28] as well as the rendered objects “Sphere”, “Turlte”, “Paperbowl”, “Queen”, and “Pumpkin” by Disney’s principled BSDF data set [29].

6.2.2 Real data sets

To effectively evaluate the robustness and performance of the presented photometric stereo methods, a better choice is to evaluate these methods on real photometric stereo images rather than synthetic images. Some data sets, such as the Gourd & Apple data set [83] and Light Stage Data Gallery [84], have been proposed for over a decade. The Gourd & Apple data set [83] consisted of three objects, namely “Apple”, “Gourd1”, and “Gourd2”, with 112, 102 and 98 images, respectively. The Light Stage Data Gallery [84] consisted of six objects and 253 images were provided for each object. However, these data sets only provided calibrated light directions without ground-truth normal maps. Therefore, one can only qualitatively compare methods on these real data sets.

To quantitatively evaluate photometric stereo methods, Shi *et al.* [6] first established a real photometric stereo data set with ground truth, namely DiLiGenT, which was the most widely used benchmark in the field of photometric stereo. This data set included ten objects with varying complexity, from simple spheres to intricate and concave geometries, and a wide range of materials, including mostly diffuse to strongly non-Lambertian surfaces with spatially varying properties. The authors illuminated and captured 96 images for each object under different lighting directions. To obtain the ground truth, the authors used a structured light-based Rexcan CS scanner, synchronized with a turn table to acquire 3D point clouds, which can calculate surface normals. Then, the shape-to-image alignment was performed to transform the 3D shape from the scanner coordinate system to the photometric stereo image coordinate system using the mutual information method in Meshlab [97]. Furthermore, the DiLiGenT benchmark [6] provided a test set, which is from a different viewpoint of these photoed objects (except for the object “Ball”) using the same lighting setup. However, using a small number of objects (10) of DiLiGenT [6] is prone to overfitting in training deep neural networks, and the shapes scanned by a 3D scanner may have errors and blurring.

To address these limitations, Ren *et al.* [80] further proposed a new real-world photometric stereo data set with ground-truth normal maps, namely DiLiGenT10^2 because it contained 10 times larger (one hundred objects of ten shapes multiplied by ten materials) than the widely used DiLiGenT benchmark [6]. The authors used ten shapes to fabricate objects, from CAD models with selected materials, through a high-precise computer numerical control (CNC) machining process, rather than scanning existing objects, which greatly avoided measurement errors. For each shape, ten materials were used to make the objects, from isotropic (diffuse and specular) and anisotropic, to translucent reflectance. Recently, Wang *et al.* [81] introduced a real-world data set, DiLiGenT-Π, for detailed near-planar surfaces. This data set was specifically designed to capture objects with high-frequency detailed structures, such as coins and badges. Similar to the DiLiGenT data set [6], the authors used a 3D scanner to acquire ground-truth 3D models for 30 objects in this data set. The presented training and test data sets for deep learning photometric stereo methods are summarized in Table 2.

7 Benchmark Evaluation Results

The evaluation metric is based on the statistics of angular errors. For the whole normal map, the mean angular error (MAE) is calculated as follows:

$$\mathrm{MAE} = \frac{1}{T}\sum_{p}^{T} \cos^{-1}(\boldsymbol{n_p}^{\top}\tilde{\boldsymbol{n}}_{\boldsymbol{p}}), \tag{10}$$

where T is the total number of pixels on the object, excluding the pixels at background positions, and $\boldsymbol{n_p}$ and $\tilde{\boldsymbol{n}}_{\boldsymbol{p}}$ are the ground-truth and estimated surface normal vector at the position indexed p. In addition to MAE, some papers also used the ratios of the number of surface normals with angular error smaller than x°, denoted as $err_{<x^\circ}$ [45], [47].

In Table 3, we report the quantitative results of the above-mentioned deep learning-based calibrated (marked as red) and uncalibrated (marked as green) photometric stereo methods on the DiLiGenT benchmark data set [6] under all the 96 input images (dense condition). Similarly, we review the performance of these calibrated deep learning-based photometric stereo methods in the sparse condition (10 input images) tabulated in Table 4. Note that not all the methods report the results under 10 input images, and some methods only provide the sparse condition without dense input, such as PS-Transformer [56].

Besides deep learning methods, we also evaluate the performance of some representative non-learning-based calibrated algorithms (marked as blue) and compare them with deep learning-based methods. As shown in Table 3, most of the learning-based methods are represented by

TABLE 3
Performance on the DiLiGenT benchmark [6] with 96 images, measured in terms of MAE in degrees. The compared methods are ranked by the average MAE of ten objects.

Method	Ball	Bear	Bear-76	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
Baseline [1]	4.10	8.39	-	14.92	8.41	25.60	18.50	30.62	8.89	14.65	19.80	15.39
IW12 [13]	2.54	7.32	-	11.11	7.21	25.70	16.25	29.26	7.74	14.09	16.17	13.74
WG10 [12]	2.06	6.50	-	10.91	6.73	25.89	15.70	30.01	7.18	13.12	15.39	13.35
HM10 [10]	3.55	11.48	-	13.05	8.40	14.95	14.89	21.79	10.85	16.37	16.82	13.22
KS21 [65]	3.78	5.96	-	13.14	7.91	10.85	11.94	25.49	8.75	10.17	18.22	11.62
IA14 [9]	3.34	7.11	-	10.47	6.74	13.05	9.71	25.95	6.64	8.77	14.19	10.60
ST14 [11]	1.74	6.12	-	10.60	6.12	13.93	10.09	25.44	6.51	8.78	13.63	10.30
SPLINE-Net† [36]	4.51	5.28	-	10.36	6.49	7.44	9.62	17.93	8.29	10.89	15.50	9.63
SDPS-Net [59]	2.77	6.89	-	8.97	8.06	8.48	11.91	17.43	8.14	7.50	14.90	9.51
DPSN [20]	2.02	6.31	-	12.68	6.54	8.01	11.28	16.86	7.05	7.86	15.51	9.41
SK22 [61]	3.46	5.48	-	10.00	8.94	6.04	9.78	17.97	7.76	7.10	15.02	9.15
IRPS [30]	1.47	5.79	-	10.36	5.44	6.32	11.47	22.59	6.09	7.76	11.03	8.83
UPS-GCNet [60]	2.50	5.60	-	8.60	7.80	8.48	9.60	16.20	7.20	7.10	14.90	8.70
LMPS [37]	2.40	5.23	-	9.89	6.11	7.98	8.61	16.18	6.54	7.48	13.68	8.41
PS-FCN [22]	2.82	7.55	5.02	7.91	6.16	7.33	8.60	15.85	7.13	7.25	13.33	8.39
ReDDLE-Net [62]	2.65	6.04	-	7.28	8.76	6.80	8.42	12.28	7.82	7.99	14.03	8.21
Manifold-PSN [50]	3.05	6.31	-	7.39	6.22	7.34	8.85	15.01	7.07	7.01	12.65	8.09
LERPS [63]	2.41	6.93	-	8.84	7.43	6.36	8.78	11.57	8.32	7.01	11.51	7.92
Attention-PSN [47]	2.93	4.86	-	7.75	6.14	6.86	8.42	15.44	6.92	6.97	12.90	7.92
DR-PSN [78]	2.27	5.46	-	7.84	5.42	7.01	8.49	15.40	7.08	7.21	12.74	7.90
GPS-Net [38]	2.92	5.07	-	7.77	5.42	6.14	9.00	15.14	6.04	7.01	13.58	7.81
JJ21 [71]	2.51	5.77	-	7.88	6.56	6.29	8.40	14.95	7.21	7.40	11.01	7.80
CHR-PSN [48]	2.26	6.35	-	7.15	5.97	6.05	8.32	15.32	7.04	6.76	12.52	7.77
CNN-PS† [21]	2.12	8.30	4.10	8.07	4.38	7.92	7.42	14.08	5.37	6.38	12.12	7.62
SPS-Net [72]	2.80	-	-	6.90	5.10	6.30	7.10	13.70	7.50	7.40	11.90	7.60
MT-PS-CNN [68]	2.29	5.79	-	6.85	5.87	7.48	7.88	13.71	6.92	6.89	11.94	7.56
HS17 [15]	1.33	5.58	-	8.48	4.88	8.23	7.57	15.81	5.16	6.41	12.08	7.55
PS-FCN (Norm.) [44]	2.67	7.72	-	7.53	4.76	6.72	7.84	12.39	6.17	7.15	10.92	7.39
MF-PSN [52]	2.07	5.83	-	6.88	5.00	5.90	7.46	13.38	7.20	6.81	12.20	7.27
HPS-Net [55]	2.37	5.28	-	6.89	4.98	5.59	7.59	14.17	6.23	6.77	11.26	7.11
LL22b [64]	1.24	3.82	-	9.28	4.72	5.53	7.12	14.96	6.73	6.50	10.54	7.05
HT21† [40]	2.49	8.96	3.59	7.23	4.69	4.89	6.89	12.79	5.10	4.98	11.08	6.91
PSMF-PSN [69]	2.54	5.99	-	7.21	5.09	5.52	7.75	11.40	6.91	6.11	10.01	6.85
NormAttention-PSN [45]	2.93	5.48	4.80	7.12	4.65	5.99	7.49	12.28	5.96	6.42	9.93	6.83
WZ20 [31]	1.78	5.26	4.12	6.09	4.66	6.33	7.22	13.34	6.46	6.45	10.05	6.76
SR-PSN [53]	2.23	5.24	-	6.75	4.63	6.12	7.07	12.61	5.88	6.44	10.35	6.73
JZ23 [79]	2.26	4.57	-	7.07	4.72	5.83	7.73	11.35	5.68	6.39	11.38	6.70
IS22† [33]	2.30	-	3.90	7.70	4.20	5.70	7.20	13.80	5.00	5.40	10.70	6.60
GR-PSN [32]	2.22	5.61	-	6.73	4.33	6.17	6.78	12.03	5.54	6.42	9.65	6.55
LL22a [77]	2.43	-	3.64	8.04	4.86	4.72	6.68	14.90	5.99	4.97	8.75	6.50
PX-Net† [39]	2.03	4.13	3.57	7.61	4.39	4.69	6.90	13.10	5.08	5.10	10.26	6.33

their networks' names. For non-learning methods and some learning-based methods without given names, we present them by the first letter of the authors' name and the published year. To ensure fairness in the evaluation, we also employ † to denote the networks trained by CyclePS [21], which are rendered using Disney's principled BSDF data set [29]. Theoretically, Disney's principled BSDFs contain an extensive range of reflectance properties by integrating various BRDFs controlled by 11 parameters. Consequently, the reflectance distributions of CyclePS more closely resemble those encountered in real-world scenarios compared to the Blobby and Sculpture data set [22], which are rendered using the MERL BRDF data set [28]. Furthermore, some recent models discarded the first 20 images of "Bear" in testing (*i.e.*, tested with the remaining 76 images) because the first 20 images are photometrically inconsistent in the belly region [21]. For these methods, we tabulate both the results of "Bear" input with 76 images and 96 images, denoted as "Bear-76" and "Bear", respectively. For a fair comparison, the average MAE of these ten objects uses the result of Bear rather than Bear-76, except for IS22 [33] and LL22a [77], which only report the Bear-76 results. Additionally, SPS-Net [72] discards the results of Bear; therefore, we can only calculate the average MAE via the remaining nine objects. Since only parallel white lights were used in the DiLiGenT benchmark [6], we can only evaluate the methods for calibrated and uncalibrated photometric stereos, ignoring methods for near-field light, general light, and color light.

Furthermore, in Figs. 10 and 11, we visualize the representative deep learning-based calibrated photometric stereo methods. The visual comparisons are based on the objects "Reading" and "Harvest' in the DiLiGenT benchmark data set [6]. For more visualization comparisons please refer to https://github.com/Kelvin-Ju/Survey-DLCPS

In Figs. 10 and 11, we evaluate the visualized reconstructed normal maps and error maps of 12 deep learning-

TABLE 4
Performance on the DiLiGenT benchmark [6] with 10 images, measured in terms of MAE in degrees. The compared methods are ranked by the average MAE of ten objects.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
IA14 [9]	12.94	16.40	20.63	15.53	18.08	18.73	32.50	6.28	14.31	24.99	19.04
Baseline [1]	5.09	11.59	16.25	9.66	27.90	19.97	33.41	11.32	18.03	19.86	17.31
ST14 [11]	5.24	9.39	15.79	9.34	26.08	19.71	30.85	9.76	15.57	20.08	16.18
IW12 [13]	3.33	7.62	13.36	8.13	25.01	18.01	29.37	8.73	14.60	16.63	14.48
CNN-PS† [21]	9.11	14.08	14.58	11.71	14.04	15.48	19.56	13.23	14.65	16.99	14.34
IRPS [30]	2.12	6.92	11.41	6.58	8.87	14.99	26.55	7.14	9.61	13.70	10.79
PS-FCN [22]	4.02	7.18	9.79	8.80	10.51	11.58	18.70	10.14	9.85	15.03	10.51
SPLINE-Net† [36]	4.96	5.99	10.07	7.52	8.80	10.43	19.05	8.77	11.79	16.13	10.35
PS-FCN (Norm.) [44]	4.38	5.92	8.98	6.30	14.66	10.96	18.04	7.05	11.91	13.23	10.14
LMPS [37]	3.97	8.73	11.36	6.69	10.19	10.46	17.33	7.30	9.74	14.37	10.02
DR-PSN [78]	3.83	7.52	9.55	7.92	9.83	10.38	17.12	9.36	9.16	14.75	9.94
CHR-PSN [48]	3.91	7.84	9.59	8.10	8.54	10.36	17.21	9.65	9.61	14.35	9.92
MT-PS-CNN [68]	4.20	8.59	8.25	7.30	10.84	10.44	16.97	8.78	9.85	13.17	9.84
JJ21 [71]	3.86	7.49	9.69	7.82	8.55	10.31	16.94	9.28	9.54	14.30	9.78
IS22 [33]	4.30	5.40	8.70	6.20	11.60	10.70	20.60	7.00	8.00	13.20	9.60
PSMF-PSN [69]	3.88	5.91	8.49	6.75	11.47	9.77	16.36	8.29	11.71	12.52	9.51
GPS-Net [38]	4.33	6.34	8.87	6.81	9.34	10.79	16.92	7.50	8.38	15.00	9.43
SPS-Net [38]	4.60	-	8.00	6.90	8.30	9.00	16.70	8.90	9.00	13.60	9.40
MF-PSN [52]	2.97	4.89	7.43	5.55	8.41	9.87	12.92	7.21	9.16	12.92	8.48
PX-Net† [39]	2.50	4.90	9.40	6.30	7.20	9.70	16.10	7.00	7.70	13.10	8.37
WJ20 [31]	2.30	5.18	7.05	5.62	7.53	8.80	15.26	7.08	8.19	10.88	7.79
PS-Transformer† [56]	3.27	4.88	8.65	5.34	6.54	9.28	14.41	6.06	6.97	11.24	7.66

based calibrated photometric stereo approaches according to our taxonomy, and the traditional least square method [1]. The baseline [1], assuming Lambertian reflectance, exhibits severe errors on specular highlights. In contrast, deep learning-based methods significantly improve results in highlight regions, demonstrating the fitting ability of deep neural networks to approximate non-Lambertian surface reflectances. As the first deep network, DPSN [20] exhibits inferior results in regions with cast shadows, such as the back of the "Reading" and the pocket of the "Harvest". This limitation arises because DPSN predicts a normal vector solely based on the reflectance observations of a single pixel, neglecting information embedded in the neighborhood of a surface point. Similar issues are observed in some methods that do not consider neighboring regions [31], [37]. PX-Net [39] incorporates global information into observation maps, leading to more accurate reconstruction results in shadow and highlight regions. However, the visualized normal map from [39] exhibits sparse noises, potentially attributed to suboptimal camera noise and self-reflection settings in the generation of global effects. On the other hand, early all-pixel methods encounter errors in regions with spatially varying reflectance [22], [47], [52], [78], such as the edge of the hat and hair. This occurs because the convolutional network processes input images in a patch-wise manner, where steep color changes impact the entire patch, such as the hat of the "Reading" and the cloth of the "Harvest". This problem is eventually addressed by the normalization operation in PS-FCN (Norm.) [44] and the double-gate normalization in NormAttention-PSN [45], which can better handle color-changed surfaces.

Furthermore, as displayed in Tables 3 and 4, the results based on deep learning-based photometric methods generally achieve better performance, as compared with non-learning methods, especially in objects with complex structure and strong non-Lambertian reflectance ("Harvest", "Reading"). This illustrates the capability and generalization of deep learning techniques. However, it can be seen that most deep learning models achieve ordinary performance on very simple objects with almost diffuse reflectance, such as "Ball". We believe that this may result from overfitting in "complex" network structures and "difficult" BRDF training data sets [28], [29] that pay more attention to non-Lambertian materials [36].

8 Future Trends

In this section, we point out some promising trends for future development, based on the discussion in the above sections. First, we focus on the problem of calibrated photometric stereo. Then, we raise the perspective of the entire photometric stereo community.

As discussed in Section 3, we compare the unique characteristics of per-pixel and all-pixel methods. These methods can be further explored and better combined. For per-pixel methods, we believe that some further developments can be found in observation maps [21], *e.g.*, how to optimize unstructured light vectors via a graph-based network [98] in the "observation map", how to embed the information from adjacent surface points in a per-pixel manner. For all-pixel methods, we believe that the fusion of inter-images (inter-patches) still needs to be improved. Existing methods applied max-pooling [22], [45], [52] or manifold learning [50] to aggregate a flexible number of input images. However, these methods are underutilized for fusing features or suffer from cumbersome training pipelines. Therefore, a better fusion strategy should be proposed, which can leverage the self-attention mechanism [27] to learn the weights of input features. Of course, a more far-sighted research direction is how to efficiently combine per-pixel and all-pixel methods, which has been initially discussed in recent combined

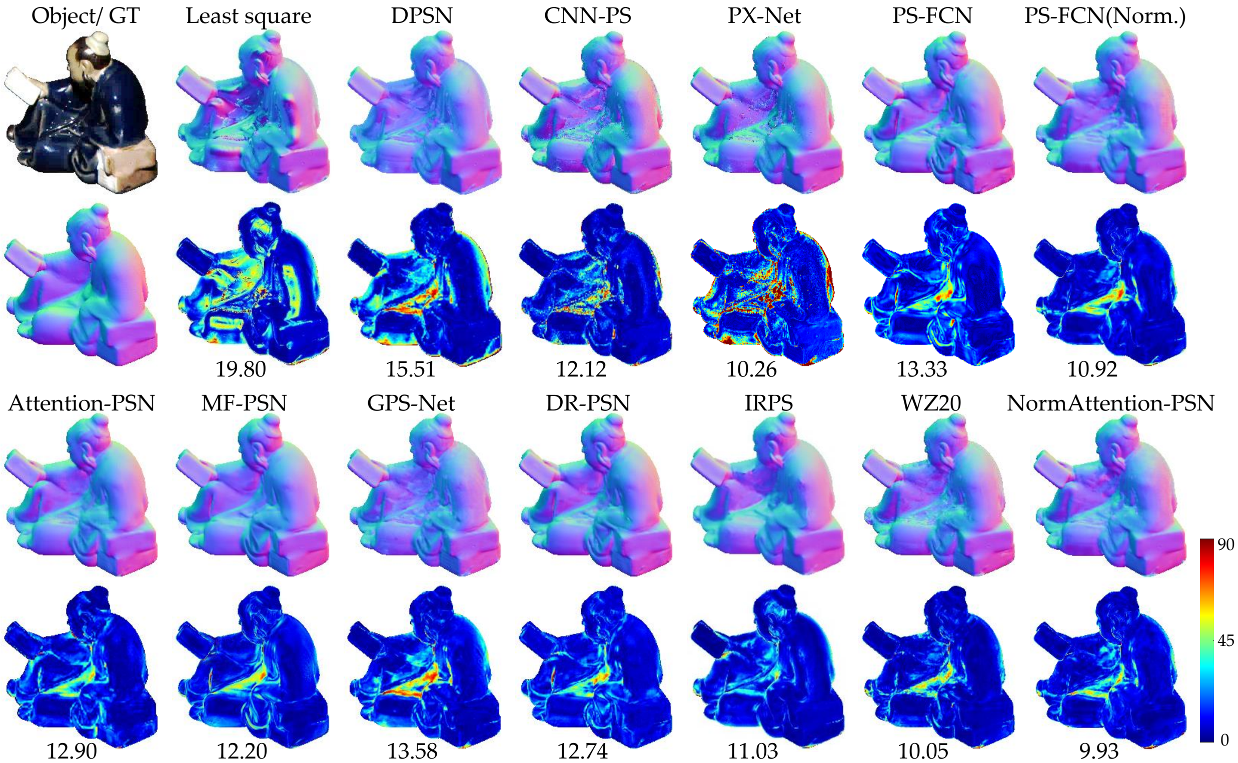


Fig. 10. Quantitative results for the object "Reading", tested with 96 input images. The first row represents the estimated normal maps, while the second row shows the corresponding error maps, with values indicating MAE in degrees.

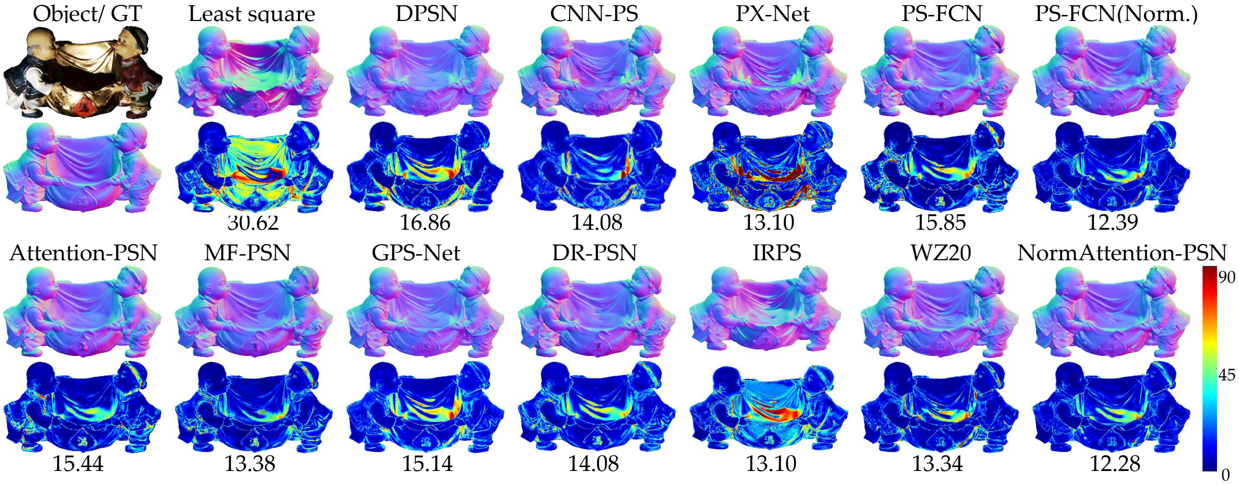


Fig. 11. Quantitative results for the object "Harvest", tested with 96 input images. The first row represents the estimated normal maps, while the second row shows the corresponding error maps, with values indicating MAE in degrees.

works [40], [68], and can be further explored by mutually combining with more physical cues. Furthermore, we argue that deep learning photometric stereo models can be further improved by excavating prior knowledge [31], [71] and supervisions [63], [67].

In fact, many deep learning-based photometric stereo methods discussed above are calibrated photometric stereo algorithms, which assume stringent requirements, such as accurate directions of incident illuminations, directional illuminations, and standard darkrooms, *etc*. In practical applications, many assumptions are not satisfied. When reviewing the realistic environment, we naturally expect a general or universal model that can handle un-calibrated light [59], [61], the colored light [67], [99], the near-field light [100], [101], the general light [102], and even a perspective projection camera simultaneously. Recently, an inspiring work UniPS [73] first dropped the physical lighting models and extracted a generic lighting representation in image interaction. This enables UniPS to accommodate various lighting scenarios, including parallel lighting, spatially varying lighting, near-field lighting, and outdoor wild lighting. Additionally, Ikehata introduced SDM-UniPS [74], designed for high-resolution input images and considering non-local interactions among surface points. SDM-UniPS [74] achieves scalable, detailed, and mask-free photometric stereo reconstruction under a universal light environment. However, these methods may face limitations when handling photometric stereo images with minor variations in lighting. This limitation arises from their reliance on the interaction mechanism for learning global lighting context rather than extracting features from each individual input. In this direction, we believe that more work can explore more effective extraction ways of universal lighting.

Furthermore, recent advancements in neural rendering technologies, *i.e.*, Neural Radiance Fields (NeRF) [103], have demonstrated great potential in photometric stereo when integrated with multi-view reconstruction. Some methods combined NeRF and multi-view photometric stereo [104], [105], which first estimate per-view surface normal maps and then blend them with a multi-view neural radiance field representation to reconstruct the object's surface geometry. Multi-view photometric stereo methods can offer a comprehensive 3D shape perception, while almost all single-view photometric stereo methods fail to recover the invisible parts (S^3-NeRF [66] can learn a neural scene representation to recover the invisible 3D parts via the single-view photometric stereo images). Notably, these NeRF-based multi-view photometric stereo techniques can avoid noticeable accumulated errors compared to traditional multi-view photometric stereo methods, which typically involve multiple disjoint and complex stages. However, existing NeRF-based photometric stereo methods still have limitations and could be explored as future trends. Firstly, NeRF-based photometric stereo methods impose a substantial computational burden and require lengthy retraining for new objects. Secondly, these methods take multi-view multi-light photometric stereo images as input, which involves fixing the camera at each viewpoint while varying the light directions. We argue that more works can explore the neural rendering techniques based on multi-view single-light, *i.e.*, light can be associated with the moving camera, potentially enhancing usability in real-world applications.

9 Conclusion

In this paper, we conducted a systematic review of deep learning-based photometric stereo methods. According to our taxonomy focusing on calibrated deep learning-based photometric stereo methods, we have summarized and discussed the strengths and weaknesses of these models by categorizing them by input processing, supervision, and network architecture. We also introduced the used training data sets and test benchmarks in the field of photometric stereo. Then, more than thirty calibrated and uncalibrated deep learning models for photometric stereo were evaluated on the widely used benchmark. Compared with traditional non-learning methods, deep learning-based photometric stereo models are superior in estimating surface normals. Finally, we pointed out the future trends in the field of photometric stereo. We hope that this survey will help researchers orient themselves to develop in this growing field, as well as highlight opportunities for future research.

Acknowledgments

The work was supported in part by the National Key R&D Program of China (2022ZD0117201), the National Key Scientific Instrument and Equipment Development Projects of China (41927805), the National Natural Science Foundation of China (62136001, 62088102, 62372306), and the Project of Strategic Importance Fund from The Hong Kong Polytechnic University (No. ZE1X).

References

[1] R. J Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 139–144, 1980.

[2] Zhenglong Zhou, Zhe Wu, and Ping Tan, "Multi-view photometric stereo with spatially varying isotropic materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1482–1489.

[3] Hao Fan, Lin Qi, Yakun Ju, Junyu Dong, and Hui Yu, "Refractive laser triangulation and photometric stereo in underwater environment," *Optical Engineering*, vol. 56, no. 11, pp. 113101, 2017.

[4] Bo Wu, Yuan Li, Wai Chung Liu, Yiran Wang, Fei Li, Yang Zhao, and He Zhang, "Centimeter-resolution topographic modeling and fine-scale analysis of craters and rocks at the chang'e-4 landing site," *Earth and Planetary Science Letters*, vol. 553, pp. 116666, 2021.

[5] Mingjun Ren, Xi Wang, Gaobo Xiao, Minghan Chen, and Lin Fu, "Fast defect inspection based on data-driven photometric stereo," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 4, pp. 1148–1156, 2018.

[6] B Shi, Z Mo, Z Wu, D Duan, SK Yeung, and P Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 271–284, 2019.

[7] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita, "Deep photometric stereo networks for determining surface normal and reflectances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 01, pp. 114–128, 2022.

[8] Steffen Herbort and Christian Wöhler, "An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods," *3D Research*, vol. 2, no. 3, pp. 4, 2011.

[9] Satoshi Ikehata and Kiyoharu Aizawa, "Photometric stereo using constrained bivariate regression for general isotropic surfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2179–2186.

[10] Tomoaki Higo, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Consensus photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1157–1164.

[11] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 6, pp. 1078–1091, 2014.

[12] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Proceedings of the Asian Conference on Computer Vision*, 2010, pp. 703–717.

[13] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa, "Robust photometric stereo using sparse regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 318–325.

[14] Tai-Pang Wu and Chi-Keung Tang, "Photometric stereo via expectation maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 546–560, 2009.

[15] Zhuo Hui and Aswin C Sankaranarayanan, "Shape and spatially-varying reflectance estimation from virtual exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 10, pp. 2060–2073, 2016.

[16] Aaron Hertzmann and Steven M Seitz, "Example-based photometric stereo: Shape reconstruction with general, varying brdfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1254–1264, 2005.

[17] Yuji Iwahori, Robert J Woodham, Hidekazu Tanaka, and Naohiro Ishii, "Neural network to reconstruct specular surface shape from its three shading images," in *Proceedings of the International Conference on Neural Networks*, 1993, vol. 2, pp. 1181–1184.

[18] Wen-Chang Cheng, "Neural-network-based photometric stereo for 3d surface reconstruction," in *Proceedings of the International Joint Conference on Neural Network*, 2006, pp. 404–410.

[19] David Elizondo, Shang-Ming Zhou, and Charalambos Chrysostomou, "Surface reconstruction techniques using neural networks to recover noisy 3d scenes," in *Proceedings of the International Conference on Artificial Neural Networks*, 2008, pp. 857–866.

[20] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita, "Deep photometric stereo network," in

Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 501–509.

[21] Satoshi Ikehata, "Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–18.

[22] Guanying Chen, Kai Han, and Kwan-Yee K Wong, "Ps-fcn: A flexible learning framework for photometric stereo," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–18.

[23] Qian Zheng, Boxin Shi, and Gang Pan, "Summary study of data-driven photometric stereo methods," *Virtual Reality and Intelligent Hardware*, vol. 2, no. 3, pp. 213–221, 2020.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, pp. 770–778.

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[26] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017.

[28] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan, "A data-driven reflectance model," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 759–769, 2003.

[29] Stephen McAuley, Stephen Hill, Naty Hoffman, Yoshiharu Gotanda, Brian Smits, Brent Burley, and Adam Martinez, "Practical physically-based shading in film and game production," in *ACM SIGGRAPH 2012 Courses*, pp. 1–7. ACM, 2012.

[30] Tatsunori Taniai and Takanori Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 4857–4866.

[31] Xi Wang, Zhenxiong Jian, and Mingjun Ren, "Non-lambertian photometric stereo network based on inverse reflectance model with collocated light," *IEEE Transactions on Image Processing*, vol. 29, pp. 6032–6042, 2020.

[32] Yakun Ju, Boxin Shi, Yang Chen, Huiyu Zhou, Junyu, and Kin-Man Lam, "Gr-psn: Learning to estimate surface normal and reconstruct photometric stereo images," *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[33] Satoshi Ikehata, "Does physical interpretability of observation map improve photometric stereo networks?," in *Proceedings of the IEEE International Conference on Image Processing*, 2022, pp. 291–295.

[34] Fotios Logothetis, Roberto Mecca, Ignas Budvytis, and Roberto Cipolla, "A cnn based approach for the point-light photometric stereo problem," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 101–120, 2023.

[35] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool, "Uncertainty-aware deep multi-view photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12601–12611.

[36] Qian Zheng, Yiming Jia, Boxin Shi, Xudong Jiang, Ling-Yu Duan, and Alex C Kot, "Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8549–8558.

[37] Junxuan Li, Antonio Robles-Kelly, Shaodi You, and Yasuyuki Matsushita, "Learning to minify photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7568–7576.

[38] Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, and Boxin Shi, "Gps-net: Graph-based photometric stereo network," in *Proceedings of Advances in Neural Information Processing Systems*, 2020, p. 33.

[39] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla, "Px-net: Simple and efficient pixel-wise training of photometric stereo networks," in *Proceedings of the International Conference on Computer Vision*, 2021, pp. 12757–12766.

[40] David Honzátko, Engin Türetken, Pascal Fua, and L Andrea Dunbar, "Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo," in *Proceedings of the International Conference on 3D Vision*, 2021, pp. 394–402.

[41] Olivia Wiles and Andrew Zisserman, "Silnet: Single-and multi-view reconstruction by learning from silhouettes," in *Proceedings of the British Machine Vision Conference*, 2017, pp. 99.1–99.13.

[42] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler, "Learned multi-patch similarity," in *Proceedings of the International Conference on Computer Vision*, 2017, pp. 1586–1594.

[43] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[44] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee Kenneth Wong, "Deep photometric stereo for non-lambertian surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 129–142, 2020.

[45] Yakun Ju, Boxin Shi, Muwei Jian, Lin Qi, Junyu Dong, and Kin-Man Lam, "Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention," *International Journal of Computer Vision*, 2022.

[46] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[47] Yakun Ju, Kin-Man Lam, Yang Chen, Lin Qi, and Junyu Dong, "Pay attention to devils: A photometric stereo network for better details," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2020, pp. 694–700.

[48] Yakun Ju, Yuxin Peng, Muwei Jian, Feng Gao, and Junyu Dong, "Learning conditional photometric stereo with high-resolution features," *Computational Visual Media*, vol. 8, no. 1, pp. 105–118, 2022.

[49] Euijeong Song and Minho Chang, "Photometric stereo using cnn-based feature-merging network," in *Proceedings of the International Conference on Control, Automation and Systems*, 2020, pp. 865–868.

[50] Yakun Ju, Muwei Jian, Junyu Dong, and Kin-Man Lam, "Learning photometric stereo via manifold-based mapping," in *2020 IEEE International Conference on Visual Communications and Image Processing*, 2020, pp. 411–414.

[51] Joshua B Tenenbaum, Vin de Silva, and John C Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[52] Yanru Liu, Yakun Ju, Muwei Jian, Feng Gao, Yuan Rao, Yeqi Hu, and Junyu Dong, "A deep-shallow and global-local multi-feature fusion network for photometric stereo," *Image and Vision Computing*, vol. 118, pp. 104368, 2022.

[53] Yakun Ju, Muwei Jian, Cong Wang, Cong Zhang, Junyu Dong, and Kin-Man Lam, "Estimating high-resolution surface normals via low-resolution photometric stereo images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, pp. 2512–2524, 2024.

[54] Clément Hardy, Yvain Quéau, and David Tschumperlé, "A multi-scale network for photometric stereo with a new comprehensive training dataset," in *Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 2023, pp. 194–203.

[55] Yakun Ju, Kin-Man Lam, Jun Xiao, Cong Zhang, Cuixin Yang, and Junyu Dong, "Efficient feature fusion for learning-based photometric stereo," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[56] Satoshi Ikehata, "Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism," in *Proceedings of the British Machine Vision Conference*, 2021, vol. 2, p. 11.

[57] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 3744–3753.

[58] Peter N Belhumeur, David J Kriegman, and Alan L Yuille, "The bas-relief ambiguity," *International Journal of Computer Vision*, vol. 35, no. 1, pp. 33–44, 1999.

[59] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong, "Self-calibrating deep photometric stereo networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8739–8747.

[60] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K Wong, and Yasuyuki Matsushita, "What is learned in deep uncalibrated photometric stereo?," in *Proceedings of the European conference on computer vision*, 2020, pp. 745–762.

[61] Francesco Sarno, Suryansh Kumar, Berk Kaya, Zhiwu Huang, Vittorio Ferrari, and Luc Van Gool, "Neural architecture search for efficient uncalibrated deep photometric stereo," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 361–371.

[62] Jiangxin Yang, Binjie Ding, Zewei He, Gang Pan, Yanpeng Cao, Yanlong Cao, and Qian Zheng, "Reddle-net: Reflectance decomposition for directional light estimation," in *Photonics*, 2022, vol. 9, p. 656.

[63] Ashish Tiwari and Shanmuganathan Raman, "Lerps: Lighting estimation and relighting for photometric stereo," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 2060–2064.

[64] Junxuan Li and Hongdong Li, "Self-calibrating photometric stereo by neural inverse rendering," in *Proceedings of the European Conference on Computer Vision*, 2022.

[65] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool, "Uncalibrated neural inverse rendering for photometric stereo of general surfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3804–3814.

[66] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong, "S^3-nerf: Neural reflectance field from shading and shadow under a single viewpoint," in *Proceedings of the Advances in Neural Information Processing Systems*, 2022, pp. 1568–1582.

[67] Yakun Ju, Xinghui Dong, Yingyu Wang, Lin Qi, and Junyu Dong, "A dual-cue network for multispectral photometric stereo," *Pattern Recognition*, vol. 100, pp. 107162, 2020.

[68] Yanlong Cao, Binjie Ding, Zewei He, Jiangxin Yang, Jingxi Chen, Yanpeng Cao, and Xin Li, "Learning inter-and intraframe representations for non-lambertian photometric stereo," *Optics and Lasers in Engineering*, vol. 150, pp. 106838, 2022.

[69] Yuze Yang, Jiahang Liu, Yue Ni, Cihan Li, and Zihan Wang, "Accurate normal measurement of non-lambertian complex surface based on photometric stereo," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[70] Jianlong Chang, Jie Gu, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "Structure-aware convolutional neural networks," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2018, pp. 11–20.

[71] Yakun Ju, Muwei Jian, Shaoxiang Guo, Yingyu Wang, Huiyu Zhou, and Junyu Dong, "Incorporating lambertian priors into surface normals measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[72] Huiyu Liu, Yunhui Yan, Kechen Song, and Han Yu, "Sps-net: Self-attention photometric stereo network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2020.

[73] Satoshi Ikehata, "Universal photometric stereo network using global lighting contexts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12591–12600.

[74] Satoshi Ikehata, "Scalable, detailed and mask-free universal photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13198–13207.

[75] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of the International Conference on Machine Learning*, 2021, vol. 2, p. 4.

[76] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 7262–7272.

[77] Junxuan Li and Hongdong Li, "Neural reflectance for shape recovery with shadow handling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16221–16230.

[78] Yakun Ju, Junyu Dong, and Sheng Chen, "Recovering surface normal and arbitrary images: A dual regression network for photometric stereo," *IEEE Transactions on Image Processing*, vol. 30, pp. 3676–3690, 2021.

[79] Yakun Ju, Cong Zhang, Songsong Huang, Yuan Rao, and Kin-Man Lam, "Learning deep photometric stereo network with reflectance priors," in *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, 2023, pp. 2027–2032.

[80] Jieji Ren, Feishi Wang, Jiahao Zhang, Qian Zheng, Mingjun Ren, and Boxin Shi, "Diligent102: A photometric stereo benchmark dataset with controlled shape and material variation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12581–12590.

[81] Feishi Wang, Jieji Ren, Heng Guo, Mingjun Ren, and Boxin Shi, "Diligent-pi: Photometric stereo for planar surfaces with rich details-benchmark dataset and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 9477–9487.

[82] Brent Burley and Walt Disney Animation Studios, "Physically-based shading at disney, part of practical physically based shading in film and game production," *Proceedings of ACM SIGGRAPH Courses*, vol. 3, pp. 5, 2012.

[83] Neil Alldrin, Todd Zickler, and David Kriegman, "Photometric stereo with non-parametric and spatially-varying reflectance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[84] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz, "Shape and spatially-varying brdfs from photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1060–1071, 2009.

[85] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs, "Large scale multi-view stereopsis evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 406–413.

[86] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al., "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[87] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 153–168, 2016.

[88] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan, "Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials," *IEEE Transactions on Image Processing*, vol. 29, pp. 4159–4173, 2020.

[89] Roberto Mecca, Fotios Logothetis, Ignas Budvytis, and Roberto Cipolla, "Luces: A dataset for near-field point light source photometric stereo," *arXiv preprint arXiv:2104.13135*, 2021.

[90] Yakun Ju, Lin Qi, Huiyu Zhou, Junyu Dong, and Liang Lu, "Demultiplexing colored images for multispectral photometric stereo via deep neural networks," *IEEE Access*, vol. 6, pp. 30804–30818, 2018.

[91] Micah K Johnson and Edward H Adelson, "Shape estimation in natural illumination," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2553–2560.

[92] Wenzel Jakob, "Mitsuba renderer," 2010.

[93] Bernardo Iraci, *Blender cycles: Lighting and rendering cookbook*, Packt Publishing Ltd, 2013.

[94] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.

[95] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky, "Neural networks for machine learning lecture 6a overview of mini–batch gradient descent," .

[96] Brian Curless and Marc Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1996, pp. 303–312.

[97] Massimiliano Corsini, Matteo Dellepiane, Federico Ponchio, and Roberto Scopigno, "Image-to-geometry registration: A mutual information method exploiting illumination-related geometric properties," in *Computer Graphics Forum*, 2009, vol. 28, pp. 1755–1764.

[98] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger, "Simplifying graph convolutional networks," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 6861–6871.

[99] Doris Antensteiner, Svorad Stolc, and Daniel Soukup, "Single image multi-spectral photometric stereo using a split u-shaped cnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[100] Hiroaki Santo, Michael Waechter, and Yasuyuki Matsushita, "Deep near-light photometric stereo for spatially varying re-

flectances," in *European Conference on Computer Vision*, 2020, pp. 137–152.

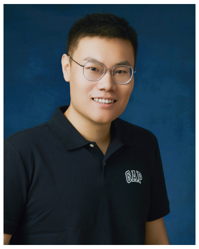
[101] Fotios Logothetis, Roberto Mecca, Ignas Budvytis, and Roberto Cipolla, "A cnn based approach for the point-light photometric stereo problem," *International Journal of Computer Vision*, pp. 1–20, 2022.

[102] Yannick Hold-Geoffroy, Paulo Gotardo, and Jean-François Lalonde, "Single day outdoor photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2062–2074, 2019.

[103] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 405–421.

[104] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool, "Neural radiance fields approach to deep multi-view photometric stereo," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 1965–1977.

[105] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong, "Ps-nerf: Neural inverse rendering for multi-view photometric stereo," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 266–284.



Yakun Ju is currently a Research Fellow of the Rapid-Rich Object Search (ROSE) Lab, School of Electrical and Electronic Engineering, Nanyang Technological University. Before that, he was a Postdoctoral Fellow of the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. He received the B.Eng. degree from Sichuan University, Chengdu, China, in 2016 and Ph.D. degree from Ocean University of China, Qingdao, China, in 2022. His research interests include computational photography, 3D reconstruction, image processing, and underwater Vision.

Kin-Man Lam received his Associateship in Electronic Engineering with distinction from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1986, his M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College, U.K., in 1987, and his Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Australia, in 1996. From 1990 to 1993, he was a lecturer at the Department of Electronic Engineering of The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University again as an Assistant Professor in 1996. He became an Associate Professor in 1999, and has been a Professor since 2010. Currently, he is also an Associate Dean of the Faculty of Engineering. He was actively involved in professional activities. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008, and was the Director-Student Services and the Director-Membership Services of the IEEE SPS between 2012 and 2014, and between 2015 and 2017, respectively. He was also the VP-Member Relations and Development and VP-Publications of the Asia-Pacific Signal and Information Processing Association (APSIPA) between 2014 and 2017, and between 2017 and 2021, respectively. He was an Associate Editor of IEEE Trans. on Image Processing between 2009 and 2014, and Digital Signal Processing between 2014 and 2018. He was also an Editor of HKIE Transactions between 2013 and 2018, and an Area Editor of the IEEE Signal Processing Magazine between 2015 and 2017. Currently, he is the IEEE SPS VP-Membership and the Member-at-Large of APSIPA. Prof. Lam also serves as a Senior Editorial Board member of APSIPA Trans. on Signal and Information Processing and an Associate Editor of EURASIP International Journal on Image and Video Processing. His current research interests include image and video processing, computer vision, and human face analysis and recognition.

Wuyuan Xie is currently an associate professor in Shenzhen University, China. Before that, she was a research associate in the Department of Computing, The Hong Kong Polytechnic University. She gained Ph.D degree in Mechanical and Automation Engineering from The Chinese University of Hong Kong in 2015. From 2008 to 2011, she was with Shenzhen Institutes of Advanced Technology, Chinese Academy of Science. Her research interests include shape from shading, motion learning, photometric stereo, and 3D surface details enhancement.

Huiyu Zhou received a Bachelor of Engineering degree in Radio Technology from Huazhong University of Science and Technology of China, and a Master of Science degree in Biomedical Engineering from University of Dundee of United Kingdom, respectively. He was awarded a Doctor of Philosophy degree in Computer Vision from Heriot-Watt University, Edinburgh, United Kingdom. Dr. Zhou currently is a full Professor at School of Computing and Mathematical Sciences, University of Leicester, United Kingdom. He has published over 500 peer-reviewed papers in the field. His research work has been or is being supported by UK EPSRC, ESRC, AHRC, MRC, EU, Royal Society, Leverhulme Trust, Invest NI, Puffin Trust, Alzheimer's Research UK, Invest NI and industry.

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 Dean of the Faculty 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.

Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the Ph.D. degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV. He is a senior member of IEEE.