




NormAttention-PSN: A High-frequency Region Enhanced Photometric Stereo Network with Normalized Attention

Yakun Ju^{1,2} · Boxin Shi^{3,4} · Muwei Jian^{5,6} · Lin Qi¹ · Junyu Dong¹  · Kin-Man Lam^{2,7}

Received: 10 July 2021 / Accepted: 30 August 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Photometric stereo aims to recover the surface normals of a 3D object from various shading cues, establishing the relationship between two-dimensional images and the object geometry. Traditional methods usually adopt simplified reflectance models to approximate the non-Lambertian surface properties, while recently, photometric stereo based on deep learning has been widely used to deal with non-Lambertian surfaces. However, previous studies are limited in dealing with high-frequency surface regions, *i.e.*, regions with rapid shape variations, such as crinkles, edges, *etc.*, resulted in blurry reconstructions. To alleviate this problem, we present a normalized attention-weighted photometric stereo network, namely NormAttention-PSN, to improve surface orientation prediction, especially for those complicated structures. In order to address these challenges, in this paper, we (1) present an attention-weighted loss to produce better surface reconstructions, which applies a higher weight to the detail-preserving gradient loss in high-frequency areas, (2) adopt a double-gate normalization method for non-Lambertian surfaces, to explicitly distinguish whether the high-frequency representation is stimulated by surface structure or spatially varying reflectance, and (3) adopt a parallel high-resolution structure to generate deep features that can maintain the high-resolution details of surface normals. Extensive experiments on public benchmark data sets show that the proposed NormAttention-PSN significantly outperforms traditional calibrated photometric stereo algorithms and state-of-the-art deep learning-based methods.

Keywords Photometric stereo · High-frequency surface normals · Non-Lambertian · Deep neural network

Communicated by Kwan-Yee Kenneth Wong.

✉ Junyu Dong
dongjunyu@ouc.edu.cn

✉ Kin-Man Lam
enkmlam@polyu.edu.hk

Yakun Ju
juyakun@stu.ouc.edu.cn

Boxin Shi
shiboxin@pku.edu.cn

Muwei Jian
jianmuwei@163.com

Lin Qi
qilin@ouc.edu.cn

¹ Department of Computer Science and Technology, Ocean University of China, Qingdao, China

² Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China

1 Introduction

Three-dimensional (3D) shape recovery is a pivotal problem in computer vision (Jian et al., 2019). Unlike binocular or multi-view stereo that use different scenes from viewpoints to triangulate sparse 3D points, photometric stereo (Woodham, 1980) recovers pixel-wise surface normals from a fixed scene under varying shading cues, which prevails in recovering fine details of the surface and dense reconstruc-

³ National Engineering Research Center of Visual Technology, School of Computer Science, and Institute for Artificial Intelligence, Peking University, Beijing, China

⁴ Peng Cheng Laboratory, Shenzhen, China

⁵ School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China

⁶ School of Information Science and Engineering, Linyi University, Linyi, China

⁷ Centre for Advances in Reliability and Safety, Tai Po, Hong Kong, China

tion. Early photometric stereo algorithms used an assumption of an ideal Lambertian reflectance (diffuse surface) model (Woodham, 1980). However, real-world objects barely have the property of Lambertian reflectance. To deal with the limitations, subsequent methods focus more on non-Lambertian surfaces with more flexible reflectance functions, adopting the bidirectional reflectance distribution functions (BRDFs) to model general reflectance (Chung & Jia, 2008; Ikehata & Aizawa, 2014; Higo et al., 2010). However, these traditional models are accurate for limited categories of materials and suffer from unstable optimization.

Meanwhile, deep learning frameworks have shown potential abilities in handling surface normal reconstruction (Wu et al., 2020; Ju et al., 2021). For photometric stereo, researchers have investigated how to learn general reflectance models through deep neural networks. DPSN (Santo et al., 2017) first addressed non-Lambertian photometric stereo using a deep fully connected network, to learn the surface normal in a per-pixel manner. Later, a series of methods employed convolutional neural networks (CNNs) to better utilize the adjacent information embedded in images, such as PS-FCN (Chen et al., 2018) and CHR-PSN (Ju et al., 2022). However, previous deep learning-based methods employed uniform cosine loss functions regardless of various surface structures and usually produce large errors in high-frequency regions, *i.e.*, regions with rapid shape variations, such as crinkles and edges. We argue that the blur and errors are caused by the following two reasons: (1) the widely used Euclidean-based loss functions hardly constrain the high-frequency representations, because

of the “regression-to-the-mean” problem (Isola et al., 2017), which results in blurry and over-smoothed images (Blau & Michaeli, 2018; Wang et al., 2004), and (2) previous learning-based photometric stereo methods usually pass the input from the high-low-high resolutions, *i.e.*, through an encoder-decoder architecture, which leads to the loss of details of prediction and causes the blur (Sun et al., 2019).

How to distinguish the high-frequency regions in an observation, caused by structure or texture, is of high importance. In fact, two conditions for identifying high-frequency representations inspire the AttentionNet: the complex surface structure (Fig. 1a) and the spatially varying BRDFs (Fig. 1b). For a complex structure, the variations of the surface normals are large, and may be considered having high-frequency representations. On the other hand, the texture of an object may cause its BRDFs to change rapidly in a flat or smooth region. Our preliminary work, Attention-PSN (Ju et al., 2020b), focuses on the structure aspect and does not consider those surfaces with spatially varying BRDFs. The high values in an attention map, in this case, may cause large errors.

In this paper, we extend Attention-PSN (Ju et al., 2020b) to handle non-Lambertian surfaces with spatially varying BRDFs, namely NormAttention-PSN, which uses our proposed double-gate observation normalization method. In Fig. 1(a), our NormAttention-PSN and Attention-PSN (Ju et al., 2020b) can produce more accurate results, but Attention-PSN cannot achieve good results in Fig. 1(b). We propose an attention-weighted loss in a self-supervised manner for NormAttention-PSN. The structure of NormAttention-PSN,

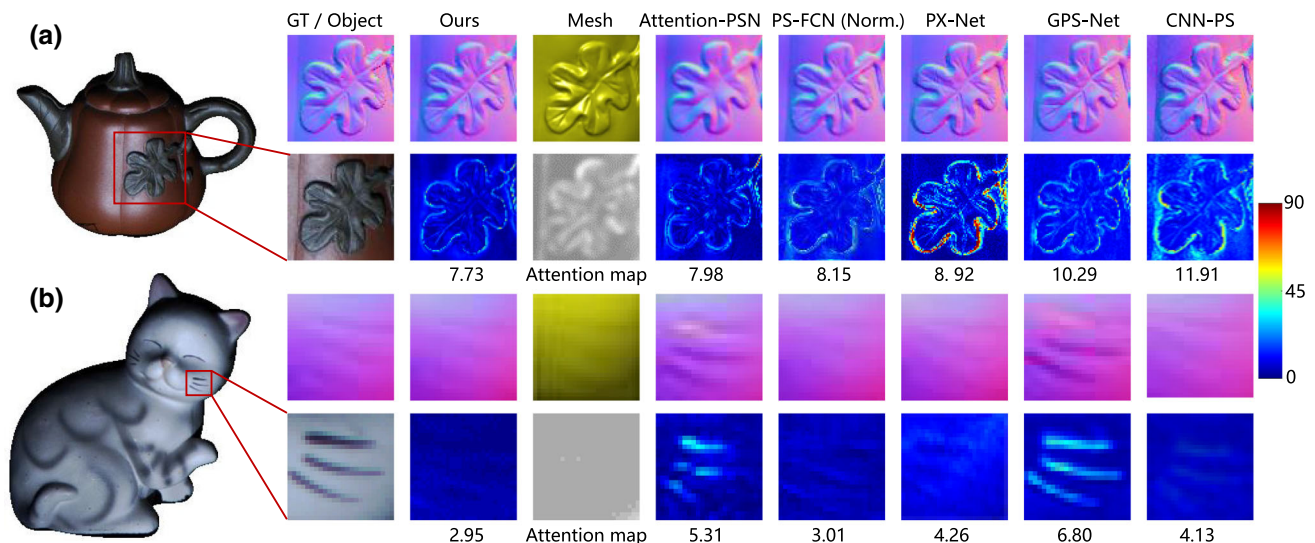


Fig. 1 Examples of the predictions and error maps on **a** high-frequency structure regions and **b** spatially varying BRDFs. We compare our method with Attention-PSN (Ju et al., 2020b), PS-FCN (Norm.) (Chen et al., 2020a), PX-Net (Logothetis et al., 2021), GPS-Net (Yao et al., 2020), and CNN-PS (Ikehata, 2018). The third column shows the 3D

reconstruction results of our estimated surface normal maps using (Simchony et al., 1990) and the generated attention maps. The numbers reveal the mean angular error in degrees. Our method obviously produces a more accurate estimation under both conditions

which is composed of two associated subnetworks, called GeometryNet and AttentionNet, respectively, is shown in Fig. 2. In short, GeometryNet reconstructs the surface normals of an object from calibrated photometric stereo images, while AttentionNet generates an attention map that provides the weights for the pixel-wise attention-weighted loss. The attention-weighted loss is composed of the angular loss and the gradient loss with adaptive weights. A pixel, which has a large value in the attention map, contains high-frequency information and should have a high detail-preserving level. Consequently, a higher weight on the gradient loss and a higher penalty on the high-frequency information should be applied.

The preliminary version of this work (Ju et al., 2020b) presents an attention-weighted loss in a self-supervised manner for each pixel, which assigns larger weights of detail-preserving penalty for high-frequency regions, to maintain the completeness of the high-frequency expression. In this paper, our explorations include three parts: (1) We extend Attention-PSN (Ju et al., 2020b) to NormAttention-PSN by adopting a double-gate observation normalization method, which can remove the impact of spatially varying BRDFs on non-Lambertian surfaces. (2) We employ a parallel high-resolution structure with multi-scale max-pooling, instead of tandem networks, inspired by the great success of the High-resolution Net (Sun et al., 2019) in human pose estimation. (3) We present a detailed network analysis and ablation experiments of each part of our method. We provide more results using both synthetic and real data sets.

Experiments have demonstrated the effectiveness of the proposed NormAttention-PSN. Our method avoids the blur in high-frequency surface regions and improves the accuracy of surface normal estimation, outperforming state-of-the-art calibrated photometric stereo approaches on public benchmark data sets.

2 Related Work

2.1 Image Formation Model and Lambertian Photometric Stereo

An imaging model for photometric stereo establishes the relationship between the 3D surface normal $\mathbf{n} \in \mathbb{R}^3$ and 2D visual observations $\{m_1, m_2, \dots, m_t\}$ in a per-pixel manner. For a pixel m_i , $i \in \{1, 2, \dots, t\}$ in the visual observation of a real-world object, viewed from direction \mathbf{v} and lighted by the illumination with direction \mathbf{l}_i with intensity e_i , $i \in \{1, 2, \dots, t\}$, the imaging model can be approximated as follows:

$$m_i = s \rho(e_i, \mathbf{n}, \mathbf{l}_i, \mathbf{v}) \max(\mathbf{n}^\top \mathbf{l}_i, 0) + \epsilon_i, \quad (1)$$

where $\rho(e_i, \mathbf{n}, \mathbf{l}_i, \mathbf{v})$ is the bidirectional reflectance distribution function (BRDF), s is a binary function for judging cast shadows ($s = 0$ for cast shadows, otherwise, $s = 1$), $\max(\mathbf{n}^\top \mathbf{l}_i, 0)$ accounts for the attached shadows, and ϵ_i reveals the noise and global illumination effect that are barely represented by the BRDF, such as inter-reflections (Nayar et al., 1991).

As an inverse problem, the goal of photometric stereo is to recover the surface orientations from a combination of reflectance and illuminations in multiple observations. Woodham (Woodham, 1980) firstly proposed a photometric stereo algorithm based on the least square method. Under the Lambertian assumption, the error term ϵ_i is ignored and BRDF has the diffuse property, where the measured intensity is proportional to the cosine of the angle between the incident light and the surface normal, but irrelevant to the viewing direction. Therefore, the imaging model can be simplified and easily solved by the least square method. This reveals that the image formation model can be cast into a system of linear equations, which can be solved. However, knowing how to make photometric stereo applicable to non-Lambertian real-world objects is more practical for the community.

2.2 Non-Lambertian Photometric Stereo

To meet the need of real-world general reflectance, researchers have investigated different strategies. Commonly referring to the taxonomy of Shi et al. (2019), we briefly divide non-Lambertian photometric stereo techniques into three categories: sophisticated reflectance methods, outlier rejection methods, and deep learning methods. More comprehensive surveys of photometric stereo can be found in Ackermann et al. (2015), Herbot and Wöhler (2011), Zheng et al. (2020).

2.2.1 Sophisticated Reflectance Methods

It is a straightforward idea to approximate real-world non-Lambertian surfaces using sophisticated reflectance models. Along this direction, researchers proposed fitting a nonlinear analytic BRDF, such as the Torrance-Sparrow model (Georghiadis, 2003), the specular spike model (Yeung et al., 2015), and the Blinn-Phong model (Tozza et al., 2016), etc. Furthermore, many studies have employed the general properties of BRDFs, such as monotonicity (Shi et al., 2012), isotropy (Higo et al., 2010), and bilateral symmetry (Alldrin & Kriegman, 2007), to deal with multiple types of surface materials. Some variants of isotropic BRDFs were proposed in Chandraker et al. (2012), Alldrin and Kriegman (2007), Holroyd et al. (2008). Shi et al. (2014) proposed a bi-polynomial model and designed an iterative strategy to represent low-frequency non-Lambertian reflectances. Ikehata and Aizawa (2014) further approximated BRDFs by using bivariate functions to deal with the instability of the

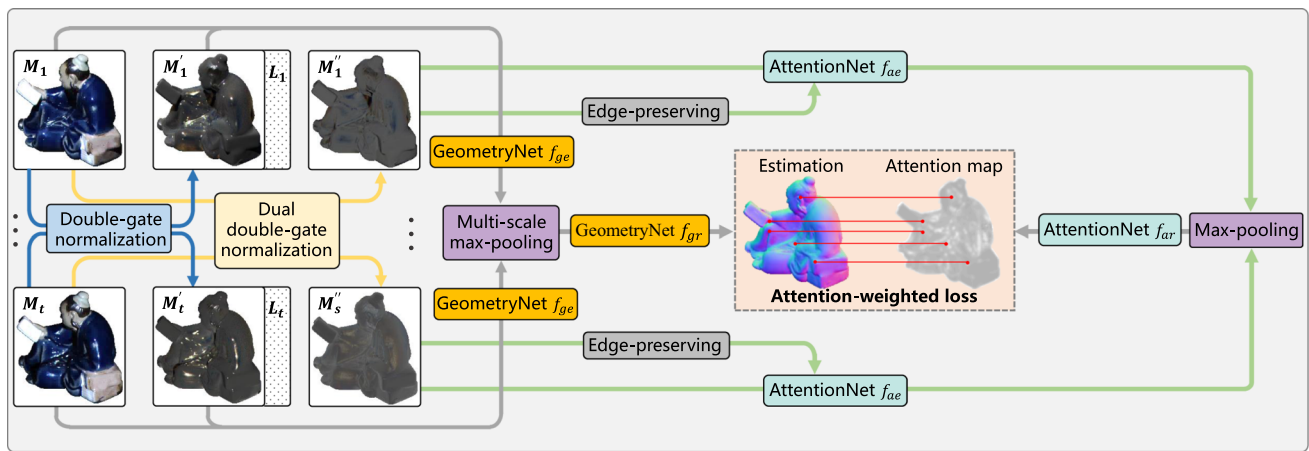


Fig. 2 Overview of our method. The attention map provides the weights for the pixel-wise attention-weighted loss. The surface normals and the corresponding attention map are learned jointly by minimizing the attention-weighted loss

estimation. However, these handcrafted analytic and empirical reflectance models are generally more useful for limited categories of reflectance, as the reflectance models vary dramatically from material to material, and these methods use numerous computations, due to the complicated solution process.

2.2.2 Outlier Rejection Methods

Outlier rejection methods assume that most regions on a surface obey, or can be approximated by, the diffuse reflectance model (Lambertian), and then those non-Lambertian regions (such as specular highlights and cast shadows) are considered as outliers, which are local and sparse. Early studies estimated surface normals by selecting three images, with the lowest specularity and the closest Lambertian appearance from multiple images (Solomon & Ikeuchi, 1996; Barsky & Petrou, 2003). Afterwards, Wu et al. (2010) proposed a robust principal component analysis (RPCA) method to decompose images into the minimized-rank Lambertian composition (Basri & Jacobs, 2003) and the non-Lambertian sparse outliers. Along this direction, Ikehata et al. (2012) employed a fixed rank of three, instead of minimizing the rank, which can achieve better computational stability. Mukaigawa et al. (2007) utilized the random sample consensus method to discard specular highlights and shadows. Furthermore, other outlier rejection methods, such as maximum-likelihood estimation (Verbiest & Van Gool, 2008), shadow cuts (Chandraker et al., 2007), maximum feasible subsystem (Yu et al., 2010), and taking the median values (Miyazaki et al., 2010), also effectively remove sparse outliers. Although robust methods are effective, broad and soft specular highlights, such as non-Lambertian diffuse reflectance, are hard to be detected as outliers. In addition, these methods usually

need a large number of observed images to achieve effective removal.

2.2.3 Deep Learning Methods

Very early studies, using neural networks, in the community of photometric stereo research can be found in (Iwahori et al., 1993; Cheng, 2006). Although the methods are effective, they require per-material pretraining or restricted Lambertian reflectance. With the recent great development of deep learning, Santo et al. Santo et al. (2017) were the first to use the modern deep neural network (DNN) architecture to predict surface normals from photometric stereo images and explore the simultaneous prediction of reflectance in Santo et al. (2020). However, the employed fully connected architecture (Santo et al., 2017) hardly benefits from the information embedded in the neighborhood of a surface point, and depends on a pre-defined set of illumination directions.

More recently, convolutional neural networks (CNNs) have been more widely introduced into the research of photometric stereo (Chen et al., 2018; Ju et al., 2020b; Taniai & Maehara, 2018; Ikehata, 2018; Ju et al., 2021; Li et al., 2019; Zheng et al., 2019; Wang et al., 2020). PS-FCN (Chen et al., 2018, 2020a; Ju et al., 2022) extracted features from a combination of observed images and illuminations, and aggregated the arbitrary features by max-pooling. Some methods (Ikehata, 2018; Li et al., 2019; Zheng et al., 2019) used another approach, called the observation map, which ranges in observation intensities, according to the light directions, to overcome the problem of requiring a fixed number of inputs. Differing from the above supervised methods, Taniai and Maehara Taniai and Maehara (2018) proposed an unsupervised method, which minimizes the reconstruction loss between original input images and the inverse

rendered images. In addition, Yao et al. Yao et al. (2020) introduced GCN (Graph Convolution Network) for learning-based photometric stereo, named GPS-Net. However, the above methods, due to employing the same learning strategy and sampling-oriented loss for all pixels on the various surfaces, fail to satisfactorily handle high-frequency regions. In contrast, our previous work (Ju et al., 2020b) proposed an adaptive attention-weighted loss to improve the performance on high-frequency areas, providing suitable penalty strategies for different surfaces. In this paper, we further extend the attention-weighted loss to handle surfaces with spatially varying BRDFs, and propose a high-resolution network structure with multi-scale max-pooling. We also show a more detailed network analysis, ablation studies, and experimental results.

3 Methodology

In this section, we present our NormAttention-PSN, which can better handle calibrated photometric stereo for high-frequency structures, such as crinkles and edges. We first introduce the proposed double-gate observation normalization method and the architecture of our framework, as shown in Fig. 2, which can be divided into two networks, GeometryNet and AttentionNet. Then, we will present the attention-weighted loss and the implementation details of our proposed framework.

3.1 Double-gate Observation Normalization

In fact, the real scenes of an object always contain spatially varying BRDFs (*e.g.*, the stripes on the back of “Cat”, as shown in Fig. 1), which are considered as high-frequency regions in our preliminary Attention-PSN model (Ju et al., 2020b). However, the corresponding surface normals in these regions of spatially varying BRDFs should be smooth, and should not be changed with the varying surface reflectance. Therefore, there is a negative impact on Attention-PSN, which assigns these regions with higher weights, resulting in a high detail-preserving loss.

To deal with this problem, we adopt a double-gate observation normalization method to remove the influence of spatially varying BRDFs and maintain the reasonable shading cues for the photometric stereo network. After the double-gate observation normalization process, our AttentionNet will only be stimulated by real high-frequency structures, such as crinkles and edges, rather than smooth regions with textures. Furthermore, another advantage is that the normalization method is beneficial to the whole training process, because photometric stereo methods are usually trained on surfaces with homogeneous BRDF and barely handle surfaces with steep color or pattern changes.

In fact, the observation normalization method was first used in PS-FCN (Norm.) (Chen et al., 2020a), which computes the normalized pixel m'_i of observations \mathbf{M}_i , $i \in \{1, 2, \dots, t\}$, as follows:

$$m'_i = \frac{m_i}{\sqrt{m_1^2 + \dots + m_t^2}}, \quad i \in \{1, 2, \dots, t\}, \quad (2)$$

where m_1, m_2, \dots, m_t are the pixel intensities of the same position in observations $\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_t$.

If the surface materials are under the Lambertian assumption, then the BRDF $\rho(e_i, \mathbf{n}, \mathbf{l}_i, \mathbf{v})$ (in Eq. 1) degenerates to a constant albedo ρ , and the observation $m_i = \rho \max(\mathbf{n}^\top \mathbf{l}_i, 0)$ (Woodham, 1980). In this case, the normalized pixel m'_i can be computed as follows:

$$m'_i = \frac{\max(\mathbf{n}^\top \mathbf{l}_i, 0)}{\sqrt{\max(\mathbf{n}^\top \mathbf{l}_1, 0)^2 + \dots + \max(\mathbf{n}^\top \mathbf{l}_t, 0)^2}}, \quad (3)$$

where the influence of albedo is removed.

However, the model $m_i = \rho \max(\mathbf{n}^\top \mathbf{l}_i, 0)$ is not applicable to non-Lambertian conditions. Although most of the regions are close to the Lambertian model, those regions with specular highlights may be impacted after the normalization process. This is due to the fact that observations without highlights will be stimulated and affected by other observations with changing specular highlights, when performing observation normalization. We visualize this problem in Fig. 3. The yellow boxes represent specular highlight regions, where the original observation normalization method (Chen et al., 2020a) cannot handle it well. Although the max-pooling operation used in the network can naturally ignore the non-activated features and only aggregate the most salient features, the suppressed observations are not equal to the suppressed features, *e.g.*, the changing appearance of an observation may cause a larger feature value (such as the yellow box of the object “Ball” by the original normalization method).

Therefore, we propose an improved method, namely double-gate observation normalization, to better handle the non-Lambertian surfaces. As illustrated in Eq. (2), a pixel under other light directions due to the specular highlights will enlarge the denominator, which causes the pixel on the normalized observation suppressed. To solve this, we propose a double-gate observation normalization method, which only uses the non-specular highlights and non-shadow observations to calculate the normalization. Concretely, inspired by the position threshold strategy (Shi et al., 2019), we set two gates, which are the corresponding positions of the lowest 10% and the highest 10% grayscale values. For a pixel m_i , if its grayscale value is between the two gates, then we retain it in the denominator of Eq. (2), otherwise, the pixel is dis-

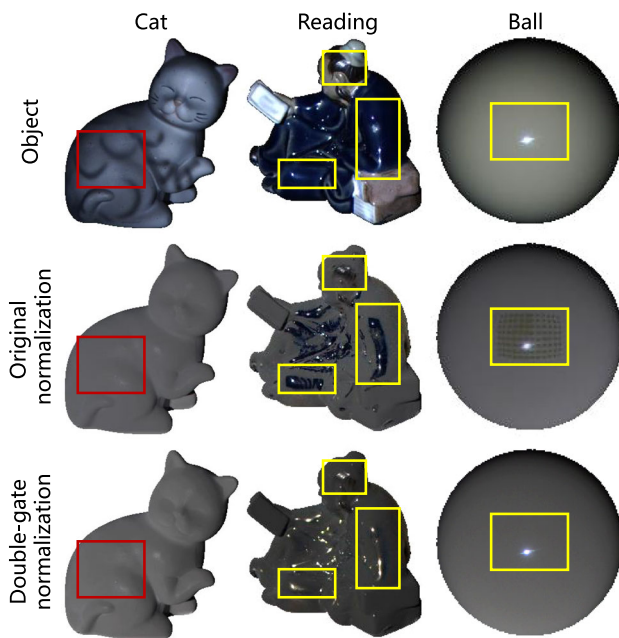


Fig. 3 Visual results based on the original normalization method (Chen et al., 2020a) and the proposed double-gate normalization method. The red boxes are regions with spatially varying BRDFs. The yellow boxes are regions with specular highlights (Color figure online)

carded and is not involved in calculating the denominator. This is because specular highlights and cast shadows often have overexposed and dark grayscale values, which can be rejected by the two gates. Therefore, the double-gate observation normalization process can be expressed as follows:

$$m'_i = \frac{m_i}{\sqrt{\sum_k m_k^2}}, \quad i \in \mathcal{T}, \quad k \in \mathcal{S}, \quad (4)$$

where the set \mathcal{S} is a subset of $\mathcal{T} = \{m_1, m_2, \dots, m_t\}$ and is controlled by the two gates, such that $m_i \in \mathcal{S}$ if $\text{Gate}(P_{10}) < m_i < \text{Gate}(P_{90})$, for $i = 1, 2, \dots, t$. The percentile P denotes a positional indicator and divides observations of all samples into two parts, where P_* means that $*\%$ of samples are smaller than it. The percentile is rounded up if it is not an integer. It is worth noting that removing some grayscale values from the denominator will affect the magnitude of the normalized observation and lead to degraded performance. We solve this problem by multiplying the normalized observation by $\sqrt{s/t}$, where s and t are the number of elements in the sets \mathcal{S} and \mathcal{T} , respectively. Furthermore, we utilize this strategy with different numbers of input images in training and testing.

As shown in Fig. 3, we compare the results of our double-gate observation normalization method with the original observation normalization method (Chen et al., 2020a). It can be seen that our method achieves more reasonable normalization results. On the regions with spatially varying BRDFs

(red boxes), such as the pattern on the back of the object “Cat”, both methods can remove the effects caused by material changes. However, on the region with specular highlights (yellow boxes), such as the arm of the object “Reading” and the middle of the object “Ball”, the position of specular highlights under other light directions will be suppressed after the previous normalization method (Chen et al., 2020a). On the contrary, our double-gate observation normalization method will avoid this condition.

3.2 Network Architecture

The proposed NormAttention-PSN is composed of two networks, GeometryNet and AttentionNet, which generate surface normals and attention maps, respectively. The details of these two networks are described below.

3.2.1 GeometryNet

GeometryNet aims to predict the surface normals \tilde{N} of an object. This network is composed of an extractor f_{ge} , a multi-scale max-pooling fusion layer, and a regressor f_{gr} , as shown in Fig. 4.

Given t normalized observations $M'_1, M'_2, \dots, M'_t \in \mathbb{R}^{3 \times H \times W}$, where $H \times W$ is the spatial resolution of the observations, and each observation has three channels, *i.e.*, RGB, we expand each illumination direction from $I_i \in \mathbb{R}^3$ to a 3-channel illumination tensor $L_i \in \mathbb{R}^{3 \times H \times W}$, having the same spatial resolution as the normalized observation image M'_i , following the previous works (Chen et al., 2018; Ju et al., 2020b; Wang et al., 2020). Here, we first concatenate the normalized observations M'_i with the original photometric stereo images M_i , forming the tensors $\Theta_i \in \mathbb{R}^{6 \times H \times W}$. We then concatenate it with the corresponding illumination directions L'_i to form the tensors $\Phi_i \in \mathbb{R}^{9 \times H \times W}$.

The reason we additionally concatenate the original images M_i is that the double-gate observation normalization may affect the shading cues for the photometric stereo network. Discarding some grayscale values in the denominator in Eq.(4) can be viewed as a kind of nonlinear processing, which influences the learning pattern of the photometric stereo network. However, the proposed double-gate observation normalization method provides more decent results on non-Lambertian surfaces, especially on regions with specular highlights. In fact, the strategy of combining the normalized images M'_i with the original images M_i will improve the accuracy of estimating surface normals.

Instead of passing the input through layers from a high to a low resolution, which are connected in series, followed by increasing the resolution, we employ a parallel, high-resolution structure for the extractor f_{ge} , inspired by the significant improvement achieved in the human pose estimation task (Sun et al., 2019). Our experiments will demonstrate

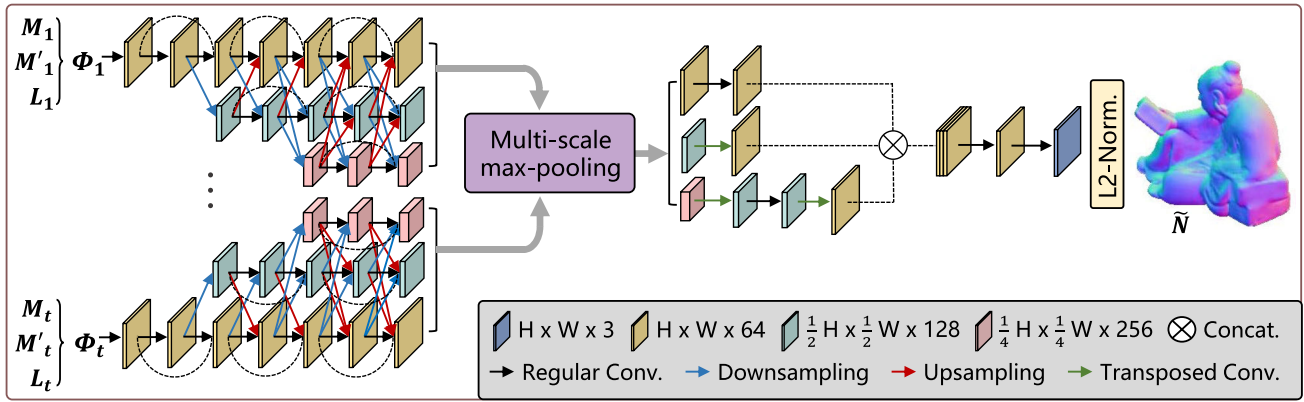


Fig. 4 The details of the high-resolution extractor f_{ge} and regressor f_{gr} of GeometryNet

that extracting high-resolution features are essential to the accuracy of the per-pixel surface normal estimation. The extractor f_{ge} can be seen as an t -multi-branch shared-weight feature extraction network, which can be expressed as follows:

$$\Psi_i^{fr}, \Psi_i^{hr}, \Psi_i^{qr} = f_{ge}(\Phi_i; \theta_{ge}), i \in \mathcal{T}, \quad (5)$$

where θ_{ge} represents the learnable parameters of the parallel high-resolution extractor f_{ge} . We employ f_{ge} to extract features at three different scales simultaneously, including the full-resolution features $\Psi_i^{fr} \in \mathbb{R}^{H \times W \times 64}$, the half-resolution features $\Psi_i^{hr} \in \mathbb{R}^{\frac{1}{2}H \times \frac{1}{2}W \times 128}$, and the quarter-resolution features $\Psi_i^{qr} \in \mathbb{R}^{\frac{1}{4}H \times \frac{1}{4}W \times 256}$, respectively (see Fig. 4). We use convolutional layers, with a stride of 2 (2 \times downsampling) or 4 (4 \times downsampling) for the downsampling operations, and adopt bilinear upsampling for the upsampling operations, with 1×1 convolutional layers to reduce the number of channels of the features. It is worth noting that the fusion of the features with different scales to form features of the same resolution is performed through skip connection rather than the concatenation operation. Furthermore, we add residual blocks (He et al., 2016) to the parallel high-resolution extractor. Residual blocks can effectively avoid vanishing gradients in deep networks, thereby further improving the accuracy of the estimated surface normals. We add the skip-connection operations to form residual blocks in each resolution branch. As shown in Fig. 4, there are three, two, and one residual blocks on the full-resolution branch, the half-resolution branch, and the quarter-resolution branch, respectively.

To enable our network to handle an arbitrary number of input observations, we apply max-pooling (Wiles & Zisserman, 2017; Chen et al., 2018; Wang et al., 2020) to fuse the t branches of multi-scale features to form a single branch. The use of max-pooling can extract the most salient information

from all the features, while average-pooling (Hartmann et al., 2017) will smooth out useful features and the generated features may be impacted by those non-activated features (Chen et al., 2018). In our method, we apply max-pooling to the three different scales to handle the t branches of multi-scale features. Denoting p as the pixel position in the extracted features, then we have

$$\left\{ \begin{array}{l} \Psi_{max}^{fr} = \bigcup_p^{H \times W} \max(\Psi_{1,p}^{fr}, \Psi_{2,p}^{fr}, \dots, \Psi_{t,p}^{fr}) \\ \Psi_{max}^{hr} = \bigcup_p^{\frac{1}{2}H \times \frac{1}{2}W} \max(\Psi_{1,p}^{hr}, \Psi_{2,p}^{hr}, \dots, \Psi_{t,p}^{hr}), \\ \Psi_{max}^{qr} = \bigcup_p^{\frac{1}{4}H \times \frac{1}{4}W} \max(\Psi_{1,p}^{qr}, \Psi_{2,p}^{qr}, \dots, \Psi_{t,p}^{qr}) \end{array} \right. \quad (6)$$

where Ψ_{max}^{fr} , Ψ_{max}^{hr} , and Ψ_{max}^{qr} are the fused features. Then, the regressor, f_{gr} with the learnable parameters θ_{gr} , takes Ψ_{max}^{fr} , Ψ_{max}^{hr} , and Ψ_{max}^{qr} as inputs and regresses the estimated surface normals \tilde{N} , as follows:

$$\tilde{N} = f_{gr}(\Psi_{max}^{fr}, \Psi_{max}^{hr}, \Psi_{max}^{qr}; \theta_{gr}). \quad (7)$$

Figure 4 shows that the resolution of the multi-scale features is adjusted by the transposed convolution operation, which upsamples the low-resolution features Ψ_{max}^{hr} and Ψ_{max}^{qr} to the full resolution of $H \times W$. In our method, concatenation is employed to fuse the two upsampled features and the full resolution feature, instead of using the skip connection in the extractor. The regressor f_{gr} uses an L2-normalization layer at the output, which makes the per-pixel estimated surface normal $\tilde{n} \in \tilde{N}$ be a unit vector.

3.2.2 AttentionNet

AttentionNet aims to generate attention maps of an object Ω . Our previous Attention-PSN (Ju et al., 2020b) fails for handling very simple structures, where the specular highlights are the only high-frequency information and therefore activate the attention map. To solve this problem, we employ an outlier rejection-based strategy for AttentionNet, to eliminate the specular highlights from the inputs of AttentionNet. Concretely, we apply dual normalization to the input images of AttentionNet. In Sect. 3.1, we only discard the smallest 10% and the highest 10% grayscale values in the denominator of Eq. (2), while retaining all grayscale values in the numerator. However, in the double-gate observation normalization, we discard the numerator and denominator (so-called dual double-gate observation normalization), as follows:

$$m''_k = \frac{m_k}{\sqrt{\sum_k m_k^2}}, k \in \mathcal{S}. \quad (8)$$

Due to discarding the numerator in the observation normalization process, we actually obtain s normalized images $M''_1, M''_2, \dots, M''_s$ from M_1, M_2, \dots, M_t . Therefore, the extractor f_{ae} of the AttentionNet only has s -multi-branch, which has fewer branches than the t -multi-branch extractor f_{ge} of the GeometryNet. In fact, the proposed dual double-gate observation normalization method does not discard the whole input images with rejected pixels, but only the pixels themselves, to retain a sufficient number of normalized images. *i.e.*, the following pixel m_{i+1} may occupy and replace the rejected pixel m_i ($i, i+1 \in \mathcal{S}$). Therefore, the normalized images may fuse multiple original photometric stereo images. Actually, this operation is fatal for photometric-based normal estimation tasks because the shading cues are disturbed. Fortunately, the AttentionNet only learns the frequency information, while paying no attention to shading cues needed by photometric stereo. Therefore, we can directly use these images normalized by the rejected double-gate observation normalization. In Fig. 5, we show an example of using different normalization methods. It can be seen that on the objects “Ball” and “Reading”, the double-gate normalized method cannot remove the highlights, and the Attention map then records all the highlights as high-frequency information, which affect the normal map estimation. However, the dual double-gate normalization removes the influence of highlights and generates a more reasonable attention map.

Similar to GeometryNet, AttentionNet is composed of an extractor f_{ae} , a fusion layer based on max-pooling, and a regressor f_{ar} . In the extractor, we first concatenate the frequency information Γ_k^{fr} with the input M'_k . The extractor f_{ae} extracts the frequency information from the k -th normalized

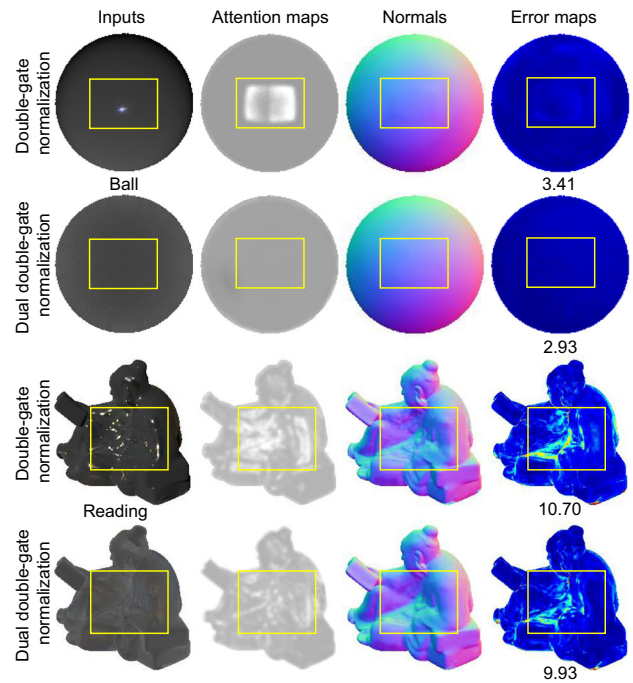


Fig. 5 The results of using double-gate normalization and dual double-gate normalization in AttentionNet. The yellow boxes represents the region with specular highlights (Color figure online)

observation, as follows:

$$\Gamma_k^{fr} = f_{ae}(M'_k, f_{ep}(M'_k); \theta_{ae}), k \in \mathcal{S}, \quad (9)$$

where f_{ae} is formed by two 3×3 convolutional layers, followed by a one-layer transposed convolutional network with learnable parameters θ_{ae} , and f_{ep} is an edge-preserving layer, which computes the gradient of M'_k . The edge-preserving layer is used to strengthen the high-frequency information of input images. We concatenate the high-frequency information $f_{ep}(M'_k)$ to the feature after the convolutional layer. Note that the output features $\Gamma_k^{fr} \in \mathbb{R}^{H \times W \times 64}$ are of full resolution. We then apply max-pooling to fuse the features of the s branches, $\Gamma_1^{fr}, \dots, \Gamma_s^{fr}$, as follows:

$$\Gamma_{max}^{fr} = \bigcup_p^{H \times W} \max(\Gamma_{1,p}^{fr}, \Gamma_{2,p}^{fr}, \dots, \Gamma_{s,p}^{fr}). \quad (10)$$

Given the fused feature Γ_{max}^{fr} , the three-layer 3×3 CNN regressor f_{ar} , with learnable parameters θ_{ar} , finally outputs the attention map Ω of the object, as follows:

$$\Omega = f_{ar}(\Gamma_{max}^{fr}; \theta_{ar}). \quad (11)$$

The attention maps $\Omega \in \mathbb{R}^{H \times W \times 1}$, provide the pixel-wise weights ω for the attention-weighted loss. In our method,

AttentionNet is learned in a self-supervised way by minimizing the attention-weighted loss, which will be introduced in Sect. 3.3.

3.3 Attention-Weighted Loss

We optimize the parameters θ_{ge} , θ_{gr} , θ_{ae} , and θ_{ar} by minimizing the attention-weighted loss, as follows:

$$\mathcal{L}_{\text{attention}} = \frac{1}{HW} \sum_p^{HW} \mathcal{L}_p, \quad (12)$$

where \mathcal{L}_p is the per-pixel loss at the pixel position p , and $H \times W$ is the resolution of the observations. \mathcal{L}_p is computed as follows:

$$\mathcal{L}_p = \lambda \omega_p \mathcal{L}_{\text{gradient}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p) + (1 - \omega_p) \mathcal{L}_{\text{angular}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p), \quad (13)$$

where ω_p is the weight obtained from the generated attention map Ω at the pixel position p . Similarly, \mathbf{n}_p and $\tilde{\mathbf{n}}_p$ are the surface-normal vector at pixel p of the ground-truth N and the estimated surface normals \tilde{N} , respectively. λ is an additional hyperparameter, which balances the angular and gradient losses and is set to 0.125, empirically (see Sect. 4.2.1), in our experiments.

$\mathcal{L}_{\text{gradient}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p)$, the first term of \mathcal{L}_p in Eq. (13), defines the gradient loss between the ground truth \mathbf{n}_p and the estimated surface normal $\tilde{\mathbf{n}}_p$, as follows:

$$\mathcal{L}_{\text{gradient}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p) = \|g(\mathbf{n}_{p(x,y)}, \xi) - g(\tilde{N}_{p(x,y)}, \xi)\|_2, \quad (14)$$

where (x, y) are the coordinates of position p . We define the gradient $g(\mathbf{n}_{p(x,y)}, \xi)$ as follows:

$$g(\mathbf{n}_{p(x,y)}, \xi) = \left\| \frac{\mathbf{n}_{p(x+\xi,y)} - \mathbf{n}_{p(x,y)}}{\xi} \right\|_1 + \left\| \frac{\mathbf{n}_{p(x,y+\xi)} - \mathbf{n}_{p(x,y)}}{\xi} \right\|_1, \quad (15)$$

where ξ is set to 1 in our setting. The gradient loss can sharpen the discontinuous or high-curvature surfaces and prevent these high-frequency regions from being blurred (Ummenhofer et al., 2017). We utilize the gradient loss to constrain the completeness and consistency of the high-frequency features. However, applying the same weight to the gradient loss, without using attention mechanisms, will result in larger errors in estimating the surface normals. This is due to the consequence of suppressing the penalty from the angular losses. Our preliminary work (Ju et al., 2020b) has proved that using only the gradient loss will cause the non-convergence

of a photometric stereo network, because gradient loss only focuses on the change between the adjacent surface normals, not on their orientations.

$\mathcal{L}_{\text{angular}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p)$, the second term in \mathcal{L}_p , is a commonly used cosine similarity loss, which directly optimizes the angular error between the ground truth \mathbf{n}_p and the estimated surface normal $\tilde{\mathbf{n}}_p$, as follows:

$$\mathcal{L}_{\text{angular}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p) = 1 - \tilde{\mathbf{n}}_p \odot \mathbf{n}_p, \quad (16)$$

where \odot represents the dot-product operation. If the predicted surface normal $\tilde{\mathbf{n}}_p$ has a similar orientation to the ground truth \mathbf{n}_p , $\tilde{\mathbf{n}}_p \odot \mathbf{n}_p$ will be close to 1 and $\mathcal{L}_{\text{angular}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p)$ will approach 0. By minimizing the above attention-weighted loss, our method learns self-supervised attention maps for different regions and brings a small angular error.

Algorithm 1 NormAttention-PSN Algorithm

Input: Photometric images M_1, M_2, \dots, M_t with illuminations I_1, I_2, \dots, I_t , hyperparameter λ .

for $j = 1$: Num_of_epochs

1. Compute the normalized M'_1, M'_2, \dots, M'_t via Eq. (4);
2. Compute the normalized $M''_1, M''_2, \dots, M''_s$ via Eq. (8);
3. Expand illuminations to L_1, L_2, \dots, L_t ;
4. Obtain \tilde{N} from $M'_1, M'_2, \dots, M'_t, M_1, M_2, \dots, M_t$, and L_1, L_2, \dots, L_t , via GeometryNet f_{ge} and f_{gr} , as shown in Eqs. (5), (6), and (7);
5. Obtain Ω from $M''_1, M''_2, \dots, M''_s$ via AttentionNet f_{ae} and f_{ar} , as shown in Eqs. (9), (10), and (11);
6. Extract ω_p and $\tilde{\mathbf{n}}_p$ from Ω and \tilde{N} , at the pixel position p ;
7. Minimize the parameters θ_{ge} , θ_{gr} , θ_{ae} , and θ_{ar} , via the attention-weighted loss Eq. (13);
8. Aggregate Ω and \tilde{N} from all ω_p and $\tilde{\mathbf{n}}_p$.

end for

Output: Estimated surface normal map \tilde{N} , attention map Ω .

Our network, with 4.63M parameters, was implemented using PyTorch. The Adam optimizer is used with the default settings ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) on a single RTX 2080 GPU. The initial learning rate is set to 0.002, and divided by 2 every 5 epochs. We trained the model using a batch size of 32, for 35 epochs. The number of input images for training is 32. In addition, we set the spatial resolution $H \times W$ to 32×32 in training. The algorithm of the proposed NormAttention-PSN is summarized in Algorithm 1.

4 Experiments

In this section, we present the experiments and analysis for our proposed framework. To evaluate the quantitative performance of our method, some widely used performance metrics are used to measure accuracy. We adopt the mean angular error (MAE) in degrees to evaluate the accuracy of the estimated surface normals, where MAE =

$\frac{1}{HW} \sum_p^{H \times W} \cos^{-1}(\mathbf{n}_p \odot \tilde{\mathbf{n}}_p)$. We also measure the ratios of the number of surface normals with angular error smaller than 10° and 30° , which are denoted as $err_{<10^\circ}$ and $err_{<30^\circ}$, respectively. $err_{<10^\circ}$ and $err_{<30^\circ}$ can better measure the errors over high-frequency regions, because the errors of surface normals in high-frequency regions are usually bigger.

4.1 Data Sets

4.1.1 Training Data Sets

For a fair comparison, two commonly used data sets, the Blobby Shape data set (Johnson & Adelson, 2011) and the Sculpture Shape data set (Wiles & Zisserman, 2017), rendered by 64 random illumination directions in the upper-hemisphere for each of the 100 BRDFs from the MERL data set (Matusik et al., 2003), were chosen to form the training set. These two data sets provide surfaces with complex structures and rich surface orientations, and the MERL dataset contains 100 different BRDFs of real-world materials. This setting has been widely used by most of the deep learning-based photometric stereo methods, such as PS-FCN (Chen et al., 2018), SDPS-Net (Chen et al., 2019), LMPS (Li et al., 2019), Attention-PSN (Ju et al., 2020b), Manifold-PSN (Ju et al., 2020a), GPS-Net (Yao et al., 2020), UPS-GCNet (Chen et al., 2020b), etc.

4.1.2 Test Data Sets

To evaluate our method, we apply several commonly used data sets, including both synthetic and real data sets. For the synthetic data set, we employ the synthetic object “Dragon” used in (Chen et al., 2020a). The object “Dragon” was rendered with 100 different BRDFs from the MERL data set (Matusik et al., 2003) under 100 random illumination directions in the upperhemisphere.

For the real data sets, we first employ the public DiLiGenT benchmark data set (Shi et al., 2019), which is composed of two parts: the main data set which contains 10 objects of various shapes with ground truth and the test data set contains 9 objects (different views from the main data set) without ground truth. Each object provides images with a resolution of 612×512 from 96 different known illumination directions. The DiLiGenT benchmark data set is challenging for its strong non-Lambertian surfaces and non-convex structures. Second, we employ the Light Stage Data Gallery (Einarsson et al., 2006) and Gourd data set (Alldrin et al., 2008), which contain six and two objects without ground truth, respectively. Each object has 253 (Light Stage Data Gallery) or 96 (Gourd data set) images under different illumination directions.

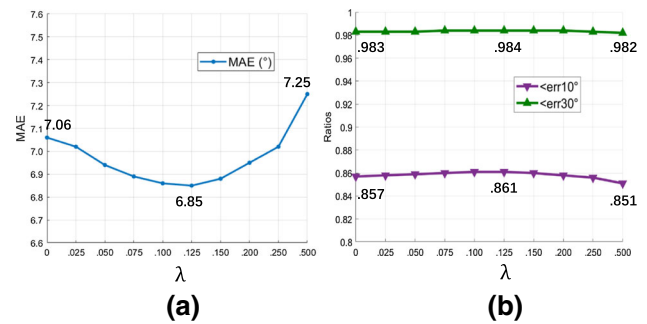


Fig. 6 Results of our NormAttention-PSN, trained with different values of the hyperparameter λ . **a** Performances in terms of MAE (the lower the better). **b** Performances in terms of $err_{<10^\circ}$ and $err_{<30^\circ}$ (the higher the better) (Color figure online)

4.2 Ablation Experiments and Network Analysis

We conducted quantitative ablation experiments on the DiLiGenT data set (Shi et al., 2019) (except for convergence comparison in Fig. 7, which uses the validation set). For all the experiments in the ablation study, we train the ablated models three times and calculate the average MAE, $err_{<10^\circ}$, and $err_{<30^\circ}$ on the DiLiGenT data set (Shi et al., 2019), with all the 96 input images.

4.2.1 Choice of the Hyperparameter λ

We first test the performances of our model with different values of the hyperparameter λ of Eq. (13). As shown in Fig. 6, it can be seen that an appropriate value of the hyperparameter λ is essential for the performance of our method. The reason why our method needs an appropriate λ can be explained as follows. The gradient loss $\mathcal{L}_{\text{gradient}}$ can highlight the high-frequency information, providing a better surface normal reconstruction in complicated regions. However, a large weight of $\mathcal{L}_{\text{gradient}}$ will dilute the penalty on the errors of the surface normal, because the gradient loss only provides the relationship between adjacent pixels, but ignores the orientation of the surface normals (as can be seen in Eqs. (14) and (15)). A more comprehensive ablated experiment about the loss function can be found in Table 1, which has also proved that using only the gradient loss will cause the non-convergence of a photometric stereo network.

To determine the optimal λ , we experimentally test our framework with different values of λ from 0 to 0.5, as shown in Fig. 6, and the best performance is achieved when $\lambda = 0.125$. With this value of λ , the average MAE is 6.85° , and the average ratios of $err_{<10^\circ}$ and $err_{<30^\circ}$ are 0.861 and 0.984, respectively, on the DiLiGenT data set (Shi et al., 2019). Note that, when the hyperparameter $\lambda = 0$, our model is trained by only using the cosine similarity loss $\mathcal{L}_{\text{angular}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p)$. In this case, the MAE is 7.06° . In fact, this performance has already outperformed PS-FCN (Norm.) (Chen et al., 2020a),

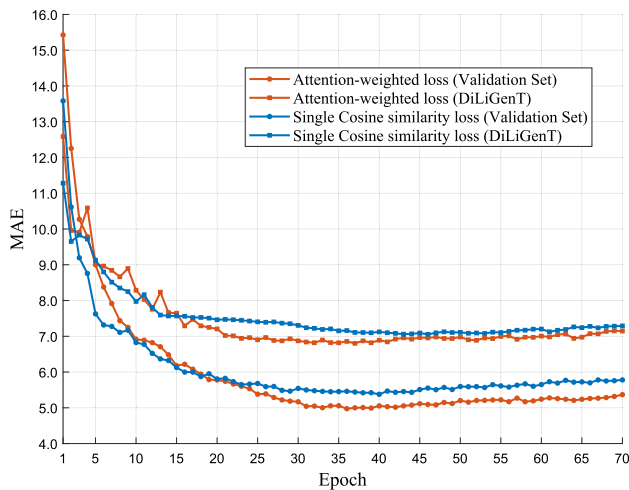


Fig. 7 Comparison of the convergence of our NormAttention-PSN on the validation set and the DiLiGenT benchmark (Shi et al., 2019). The orange line represents the results of NormAttention-PSN with the attention-weighted loss, while the blue line is the loss of GeometryNet optimized by a single cosine similarity loss. Both the networks were trained with the same structure (Color figure online)

Table 1 Ablation results for the different loss functions

Loss function	MAE ↓	$err_{<10^\circ}$ ↑	$err_{<30^\circ}$ ↑
Attention-weighted loss	6.85	0.861	0.984
1 (Cosine): 1 (Gradient)	7.74	0.849	0.980
2 (Cosine): 1 (Gradient)	7.31	0.850	0.982
5 (Cosine): 1 (Gradient)	7.23	0.852	0.982
Single cosine loss	7.06	0.857	0.983
Single gradient loss	31.56	0.156	0.595

Bold values indicate the best performance

which achieves the MAE of 7.39° on the DiLiGenT data set (Shi et al., 2019), under 96 input images. This also proves the effectiveness of our parallel high-resolution extractor f_{ge} (see Sect. 4.2.3) and double-gate normalization method (see Sect. 4.2.4).

4.2.2 Effectiveness of the Attention-Weighted Loss

In this Section, we first discuss the optimization and effectiveness of the attention-weighted loss. In Fig. 7, we visualize the convergence of our proposed model on the validation set and the DiLiGenT benchmark (Shi et al., 2019) during training our NormAttention-PSN with attention-weighted loss (orange line), which is compared to the same GeometryNet with a single cosine similarity loss $\mathcal{L}_{\text{angular}}(\mathbf{n}_p, \tilde{\mathbf{n}}_p)$ (blue line). Following the previous settings (Chen et al., 2018, 2020a), the validation set is randomly split from the training dataset, and has a total of 852 samples with 32 images. In Fig. 7, we report the average MAE of these 852 samples from the Validation set, as well as that of 96 input images from 10 objects in DiLiGenT.

As shown in Fig. 7, our NormAttention-PSN with the attention-weighted loss can achieve lower convergence error than the single cosine loss on the validation set and the DiLiGenT benchmark (Shi et al., 2019) (evaluated on Epoch 35). This illustrates the effectiveness of the attention-weighted loss. However, we also found that the attention-weighted loss has a slower speed of convergence at the beginning of training, compared with the single cosine loss. This might be explained by the fact that the generated attention maps provide inaccurate weights at the beginning of the training period.

To explicitly show the evolution of the attention maps generated by AttentionNet, we further visualize the attention maps in different training periods, as shown in Fig. 8. We show the attention maps for providing the weight ω_p of Eq. (13) from the beginning to the end of the learning period, and the error maps.

As shown in Fig. 8, the attention maps can quickly reflect the high-frequency representations. Fig. 8 shows some examples of the weights of the gradient loss (ω_p in Eq. (13)) in high-frequency regions. It can be seen that the learned weights enlarge gradually along with training, so the weight of the gradient loss $\mathcal{L}_{\text{gradient}}$ becomes more significant. In fact, the attention maps basically provide reasonable weights for the attention-weighted loss after training for only two epochs. This can be explained by the fact that the edge-preserving layer f_{ep} is fused in the AttentionNet. The edge-preserving layer computes the gradient of the normalized images \mathbf{M}'_k , which provides the obviously prior information for the edge and crinkle regions. Therefore, our AttentionNet can learn the accurate attention maps fast. In addition, it can be seen that the belly of the object “Buddha” has a region of specular highlights, whereas the attention maps in the region are not activated. It illustrates the effectiveness of our dual double-gate normalization method used in AttentionNet.

As tabulated in Table 1, we then compare the results based on different loss functions, conducted on the DiLiGenT benchmark (Shi et al., 2019). In Table 1, we evaluate the attention-weighted loss, fixed rate of cosine and gradient loss, single cosine loss, and single gradient loss. For the experiment without attention-weighted loss, we only remain the GeometryNet without the AttentionNet.

As tabulated in Table 1, the attention-weighted loss consistently outperforms the others in all the metrics. For the attention-weighted loss, a higher err_{10° and err_{30° mean that fewer complex-structured regions suffer from the large angular error. In addition, it can be seen that all the fixed combined losses achieves worse results on the metrics of MAE, $err_{<10^\circ}$, and $err_{<30^\circ}$. As discussed in Sect. 4.2.1, the gradient loss only provides the relationship between adjacent pixels, but ignores the orientation of the surface normals. The fixed combination of cosine loss and gradient loss actually

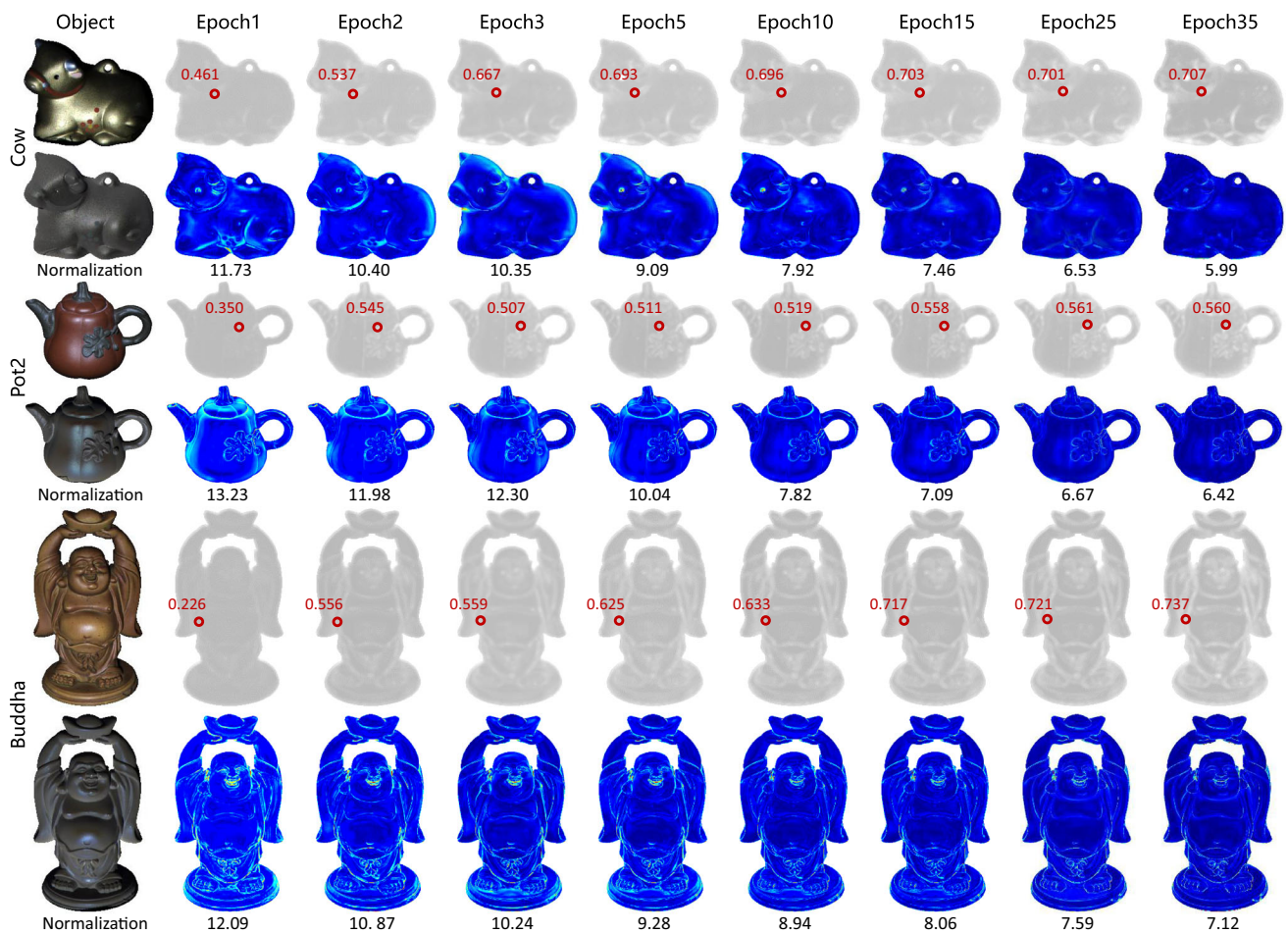


Fig. 8 The evolution of the attention maps, error maps, and the learned weights of the gradient loss in the high-frequency regions. We show the results using the checkpoints of Epoch1, Epoch2, Epoch3, Epoch5, Epoch10, Epoch15, Epoch25, and Epoch35, taking objects “Cow”,

“Pot2”, and “Buddha” from DiLiGenT benchmark (Shi et al., 2019) as the examples. The black numbers under error maps represent the MAE, while the red numbers represent the learned weights

dilute the penalty on the errors of the surface normals. It can also be proved by the ablation experiment with a single gradient loss, which cannot optimize GeometryNet. Therefore, we use an adaptive learned attention-weighted loss to optimize our framework, which only provides higher weights on the high-frequency regions to maintain the completeness of the complex structure.

4.2.3 Effectiveness of the High-Resolution Extractor

We evaluate the performance of the high-resolution structure f_{ge} . Table 2 shows the results, where ID (0) represents the performance of our proposed model, when the features are of full resolution (fr), half resolution (hr), and quarter resolution (qr), with residual blocks, as shown in Fig. 4. ID (1) compares the results of the plain layers counterpart (same high-resolution structure without residual blocks) of ID (0). For IDs (2) ~ (6), we adjust the architecture of GeometryNet to realize different combinations of the features of

Table 2 Ablated results of the parallel high-resolution structure f_{ge}

ID	Method	MAE ↓	$err_{<10^\circ}$ ↑	$err_{<30^\circ}$ ↑
(0)	Ours	6.85	0.861	0.984
(1)	<i>w/o</i> Residual blocks	6.99	0.856	0.982
(2)	<i>fr</i>	7.28	0.848	0.980
(3)	<i>fr+hr</i>	7.07	0.852	0.982
(4)	<i>fr+qr</i>	7.05	0.853	0.982
(5)	<i>hr+qr</i>	7.23	0.850	0.981
(6)	<i>fr+hr+qr+er</i>	6.95	0.858	0.982

Bold values indicate the best performance

different resolutions (without residual blocks). Note that er in ID (6) means using features of one-eighth resolution. Furthermore, no multi-scale max-pooling operation is used in ID (2), because features of a single resolution are extracted in this ablated method.

Table 3 Results of our model with different pre-processing methods

Method	MAE ↓	$err_{<10^\circ}$ ↑	$err_{<30^\circ}$ ↑
Double-gate normalization	6.85	0.861	0.984
Original normalization	6.98	0.855	0.983
<i>w/o</i> Normalization	7.56	0.848	0.980
<i>w/o</i> Dual double-gate	6.88	0.861	0.984

Bold values indicate the best performance

Table 2 tabulates the results of the ablation experiments with different structures of GeometryNet f_{ge} . Experiments (0) and (1) show the effectiveness of the residual blocks (He et al., 2016), where all the performances metrics are worse, when the residual blocks of high-resolution structure are replaced by the plain convolutional layers. Note that ID (2) has only full resolution feature, which can be seen as a fully convolutional network without up and down sampling. Referring to the experiment results for IDs (3) ~ (6), we compare the performance of using different combinations of feature resolutions. In all, combining features with multiple resolutions is beneficial to the prediction accuracy. From the experiment results for IDs (3), (4), and (5), it can be seen that combining features with higher resolutions can provide better performance. Especially, when the network has not full resolution features fr (ID (5)), the performance is worse than ID (4) with fr . It illustrates that the high resolution of features has a crucial impact on the performance of the per-pixel surface-normal recovery task. We also found that the default structure ($fr + hr + qr$, ID (1)) is slightly worse than ID (6) that has an additional resolution feature er . However, the additional er resolution feature significantly increases the parameters and training time.

4.2.4 Effectiveness of the Double-Gate Normalization

We evaluate the effectiveness of the proposed double-gate normalization, and the results are tabulated in Table 3. The results in Table 3 show the performance of our model with normalization or without normalization. Concretely, we compare the double-gate normalization with the original normalization (Chen et al., 2020a), and without the normalization method. Specially, we also report the results of without using the proposed dual double-gate normalization in AttentionNet.

As reported in Table 3, the proposed double-gate normalization method outperforms the original normalization (Chen et al., 2020a). The original method suffers from suppressed observations due to specular highlights existing on the non-Lambertian surfaces, which influences the results of the estimated surface normals. Note that our double-gate normalization method is additionally fused with the original observations M_1, M_2, \dots, M_t , as discussed in Sect. 3.2.1.

Table 4 Ablation results for different fusion methods

Fusion type	MAE ↓	$err_{<10^\circ}$ ↑	$err_{<30^\circ}$ ↑
Max-pooling	6.85	0.861	0.984
Average-pooling	7.84	0.818	0.976
Max-p. + Average-p.	7.04	0.855	0.983

Bold values indicate the best performance

The network without any observation normalization leads to a quite large angular error. With the normalization pre-processing step, the generated attention map can accurately reflect the regions with real high-frequency structures rather than being stimulated by spatially varying BRDFs.

In addition, the dual double-gate normalization used in AttentionNet can slightly improve the accuracy of the estimation. This is because the dual double-gate normalization can further remove outliers (specular highlights, *etc.*) in the observations, which might impact the generated attention maps. As shown in Fig. 5, the middle of the object “Ball” is not a complex structure region, but only the proposed dual double-gate normalization obviously can generate more reasonable attention map for the object. However, the proposed method only improves the results on the objects with particularly simple structure, where the specular highlights are the only high-frequency information. Therefore, the improvement is not very obvious.

4.2.5 Performances of the Different Fusion Methods

We further compare the performance of our framework based on different fusion methods, to experimentally explore the effect of the maximum and average operations in Table 4. We test our method with different methods, including fusion using max-pooling only, average-pooling only, and a combination of max-pooling and average-pooling (by concatenation and a 1×1 convolutional layer to keep the same number of channels). Note that the fusion layer in GeometryNet and AttentionNet are the same in our ablation experiments.

From Table 4, we can see that the max-pooling operation achieves the best performance on all three metrics. Our experimental results show a contrary conclusion, when compared to (Yao et al., 2020), which reported that a better performance can be achieved based on a combination of max-pooling and average-pooling. We conjecture that this may be due to the use of observation normalization in our method, because this operation can be partially viewed as “average-pooling” (the normalized m'_i contains information from all the original observations, see Eq. (2)). Therefore, the average-pooling operation in the fusion layer may result in redundancy in the features.

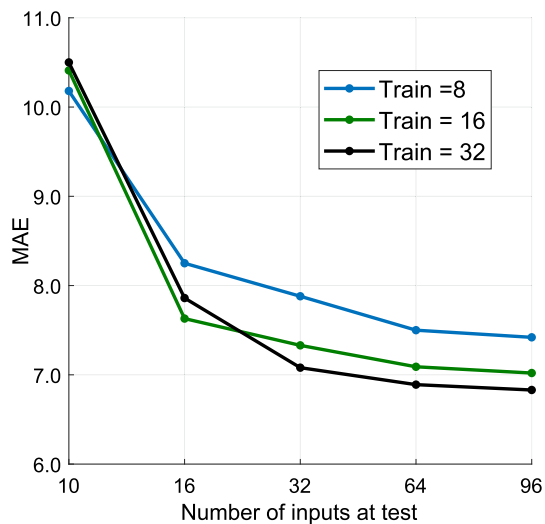


Fig. 9 Results of NormAttention-PSN trained and tested with different numbers of input images

4.2.6 Different Numbers of Inputs During Training

In this Section, we show how the number of inputs, used in training, influences the performance. Fig. 9 shows the test performance of our model, on the DiLiGenT data set (Shi et al., 2019), with different numbers of input training images. The results show that the test performance is the best, when the same number of inputs are used for training and testing, *i.e.*, NormAttention-PSN performs better, when the number of input images for training is close to that for testing. This suggests that the performance of NormAttention-PSN can be further improved by using a close number of input images for training and testing, if the number of input images for testing is known and fixed.

4.3 Benchmark Comparisons

4.3.1 DiLiGenT benchmark main data set

The test results of our method and other state-of-the-art methods on the DiLiGenT benchmark main data set (Shi et al., 2019), with 96 input images, are listed in Table 5. We compare NormAttention-PSN with traditional methods (represented by the first letter of the authors' name + published year) and learning-based (represented by their networks' names) methods in terms of MAE.

Table 5 tabulates the experimental results, in terms of MAE, for different methods on the DiLiGenT main data set (Shi et al., 2019) with all the 96 input images. It can be seen that our NormAttention-PSN achieves superior results among more than twenty methods. Compared with the same training dataset (*i.e.*, the MERL reflectance dataset (Matusik et al., 2003)), the proposed NormAttention-PSN outperforms

all the deep learning-based methods, even for the Inverse model (Wang et al., 2020) with an additional collocated light image. However, the average MAE of NormAttention-PSN on the DiLiGenT benchmark (Shi et al., 2019) is slightly worse than the two new methods LSPC-Net (Honzátko et al., 2021) and PX-Net (Logothetis et al., 2021). Nevertheless, NormAttention-PSN outperforms them on the strongly non-Lambertian objects with complex structures, such as “Buddha”, “Harvest”, and “Reading”, as shown by the visual results in Fig. 10. It can be seen that NormAttention-PSN can more accurately recover the surface normals in those regions with cast shadows, such as the sack of “Harvest” and the middle of “Reading”. In fact, LSPC-Net is trained with the CyclePS dataset (Ikehata, 2018), which is rendered by Disney’s principled BSDFs (McAuley et al., 2012). Theoretically, the Disney’s principled BSDFs used contains unlimited reflectance, since they integrates different BRDFs controlled by 11 parameters. This makes the reflectance distributions more similar in the real-world scenarios. PX-Net is further trained by a private synthetic training dataset, rendered by the BSDFs. Conversely, the MERL BRDFs dataset only contains 100 kinds of reflectance, which barely span the whole set of materials existing in nature. However, the CyclePS dataset (Ikehata, 2018) is inappropriate for most deep-learning methods since it is designed for per-pixel processing strategy rather than the all-pixel networks (Yao et al., 2020; Ju et al., 2021).

Furthermore, we also show the estimations of the high-frequency regions. In Fig. 1(a), the “leaf” of the object “Pot2” is a region with complex structure. It can be seen that the attention map of our method is activated, which reflects the frequency of the structure. With the attention-weighted loss, NormAttention-PSN outperforms all the other deep learning-based photometric stereo methods in this region. The error map of NormAttention-PSN clearly shows the less angular error of the edge and pattern of the “leaf”. Referring to Fig. 1(b), our method also avoids the influence of the steep changed materials on the surfaces, where our attention map is not activated by the spatially varying BRDFs. This illustrates the effectiveness of the proposed normalization method in AttentionNet.

In addition, as indicated in the footnote of Table 5, the original CNN-PS (Ikehata, 2018), Inverse model (Wang et al., 2020), LSPC-Net (Honzátko et al., 2021), and PX-Net (Logothetis et al., 2021) discard 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 (Ikehata, 2018). In fact, when discarding the first 20 images, the results of our NormAttention-PSN even performs 4.65 on “Bear” and 6.50 on average. In Fig. 11, we compare the results of our NormAttention-PSN, CNN-PS, PX-Net, and Inverse model when using all the 76 input images or 96 input images. It can be seen that

Table 5 Comparison of different methods on the DiLiGenT benchmark main data set (Shi et al., 2019)

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Average
Least square (Woodham, 1980)	4.10	8.39	14.92	8.41	25.60	18.50	30.62	8.89	14.65	19.80	15.39
ST12 (Shi et al., 2012)	13.58	19.44	18.37	12.34	7.62	17.80	19.30	10.37	9.84	17.17	14.58
IW12 (Ikehata et al., 2012)	2.54	7.32	11.11	7.21	25.70	16.25	29.26	7.74	14.09	16.17	13.74
WG10 (Wu et al., 2010)	2.06	6.50	10.91	6.73	25.89	15.70	30.01	7.18	13.12	15.39	13.35
HM10 (Higo et al., 2010)	3.55	11.48	13.05	8.40	14.95	14.89	21.79	10.85	16.37	16.82	13.22
AZ08 (Alldrin et al., 2008)	2.71	5.96	12.54	6.53	21.48	13.93	30.50	7.23	11.03	14.17	12.61
GC10 (Goldman et al., 2010)	3.21	6.62	14.85	8.22	9.55	14.22	27.84	8.53	7.90	19.07	12.00
IA14 (Ikehata & Aizawa, 2014)	3.34	7.11	10.47	6.74	13.05	9.71	25.95	6.64	8.77	14.19	10.60
ST14 (Shi et al., 2014)	1.74	6.12	10.60	6.12	13.93	10.09	25.44	6.51	8.78	13.63	10.30
HS17 (Hui & Sankaranarayanan, 2016))	1.33	5.58	8.48	4.88	8.23	7.57	15.81	5.16	6.41	12.08	7.55
DPSN (Santo et al., 2017)	2.02	6.31	12.68	6.54	8.01	11.28	16.86	7.05	7.86	15.51	9.41
IRPS (Taniai & Maehara, 2018)	1.47	5.79	10.36	5.44	6.32	11.47	22.59	6.09	7.76	11.03	8.83
LMPS (Li et al., 2019)	2.40	5.23	9.89	6.11	7.98	8.61	16.18	6.54	7.48	13.68	8.41
PS-FCN (Chen et al., 2018)	2.82	7.55	7.91	6.16	7.33	8.60	15.85	7.13	7.25	13.33	8.39
Manifold-PSN (Ju et al., 2020a)	3.05	6.31	7.39	6.22	7.34	8.85	15.01	7.07	7.01	12.65	8.09
Attention-PSN (Ju et al., 2020b)	2.93	4.86	7.75	6.14	6.86	8.42	15.44	6.92	6.97	12.90	7.92
DR-PSN (Ju et al., 2021)	2.27	5.46	7.84	5.42	7.01	8.49	15.40	7.08	7.21	12.74	7.90
GPS-Net (Yao et al., 2020)	2.92	5.07	7.77	5.42	6.14	9.00	15.14	6.04	7.01	13.58	7.81
CHR-PSN (Ju et al., 2022)	2.26	6.35	7.15	5.97	6.05	8.32	15.32	7.04	6.76	12.52	7.77
PS-FCN (Norm.) (Chen et al., 2020a)	2.67	7.72	7.53	4.76	6.72	7.84	12.39	6.17	7.15	10.92	7.39
Inverse model*† (Wang et al., 2020)	1.78	5.26	6.09	4.66	6.33	7.22	13.34	6.46	6.45	10.05	6.76
SPLINE-Net ‡ (Zheng et al., 2019)	4.51	5.28	10.36	6.49	7.44	9.62	17.93	8.29	10.89	15.50	9.63
CNN-PS* ‡ (Ikehata, 2018)	2.12	8.30	8.07	4.38	7.92	7.42	14.08	5.37	6.38	12.12	7.62
LSPC-Net* ‡ (Honzátko et al., 2021)	2.49	8.96	7.23	4.69	4.89	6.89	12.79	5.10	4.98	11.08	6.91
PX-Net* ‡' (Logothetis et al., 2021)	2.03	4.13	7.61	4.39	4.69	6.90	13.10	5.08	5.10	10.26	6.33
NormAttention-PSN (ours)	2.93	5.48	7.12	4.65	5.99	7.49	12.28	5.96	6.42	9.93	6.83

For those methods with *, we report the results of the object “Bear” with all the 96 input images. In their original paper, they only test the results of “Bear” with 76 input images, with the first 20 images discarded. The Inverse model (Wang et al., 2020) with † uses a specific collocated illumination constraint, so the comparison is not fully fair. The three methods with ‡ indicate that the training dataset is the CyclePS rendered by Disney’s principled BSDFs (McAuley et al., 2012), the PX-Net with ‡' indicates that the private synthetic training dataset rendered by BSDFs is used, while the other compared deep learning-based methods are all trained with the dataset rendered by the MERL reflectance dataset (Matusik et al., 2003)

the MAE of our method increases slightly when adding the first 20 “photometrically inconsistent” images, while others becomes largely worse. When adding the all 96 images of the object “Bear”, the angular error of our NormAttention-PSN only increase 14.17% (4.80 → 5.48), while it increases by 15.69% (3.57 → 4.13), 28.70% (4.12 → 5.26), 97.62% (4.20 → 8.30), and 150.28% (3.58 → 8.96) of PX-Net, Inverse model, CNN-PS, and LSPC-Net, respectively. It illustrates the robustness of the proposed method when meeting wrong illuminations.

In fact, many practical applications involve sparse photometric stereo. We evaluate our NormAttention-PSN with different numbers of input images, and compare it with recent deep learning-based methods, such as PS-FCN (Norm.) (Chen et al., 2020a), GPS-Net (Yao et al., 2020), CNN-PS (Ikehata, 2018), PS-FCN (Chen et al., 2018), IRPS (Taniai & Maehara, 2018), LMPS (Li et al., 2019), and SPLINE-Net

(Zheng et al., 2019), on the DiLiGenT main data set (Shi et al., 2019). Fig. 12 shows the comparison results.

We can see that our NormAttention-PSN outperforms all the other methods, when more than 16 images are used as inputs, and keep the promising performance when 10 input images are used. It is worth noting that some methods are trained with only 10 input images, such as LMPS (Li et al., 2019) and SPLINE-Net (Zheng et al., 2019), while our method is trained with 32 input images. When our method is also trained with fewer input images (*e.g.*, 8 and 16), the results of testing with 8 inputs are much better than those trained with 32 input images, as illustrated in Sect. 4.2.6. Furthermore, it can be seen that the errors of CNN-PS (Ikehata, 2018) is even slightly higher, when the number of input images increases from 64 to 96, because of the large errors of inputting all 96 images of the object “Bear” (see Fig. 11).



Fig. 10 Quantitative results on objects “Harvest”, “Reading”, and “Buddha” on the DiLiGenT benchmark data set (Shi et al., 2019), with 96 input images. The third column shows the 3D reconstruction results of our estimated surface normal maps using (Simchony et al., 1990) and

the generated attention maps. Compared with PX-Net (Logothetis et al., 2021), PS-FCN (Norm.) (Chen et al., 2020a), Inverse model (Wang et al., 2020), CNN-PS (Ikehata, 2018), and GPS-Net (Yao et al., 2020), our NormAttention-PSN achieves the best or sub-optimal results

4.3.2 DiLiGenT Benchmark Test Data Set

We further evaluated our model on the test data set of the DiLiGenT benchmark (Shi et al., 2019), which contains 9 objects with different views from the main data set. Due to the fact that the ground truths are not open, we can only perform limited comparisons.¹ Table 6 tabulates the results. Similar

to the results on the main data set, NormAttentionPSN outperforms other methods on this test data set. Our method achieves either the best or the second-best results on most objects. Moreover, our method even outperforms the second-best method, *i.e.*, PS-FCN (Norm.), by more than 0.55° , in terms of the average MAE, which is 0.36° (as shown in Table 5) based on the DiLiGenT main data set. The visual examples are illustrated in Fig. 13.

¹ We thank Dr. Zhipeng Mo for helping us test the results on the DiLiGenT test data set.

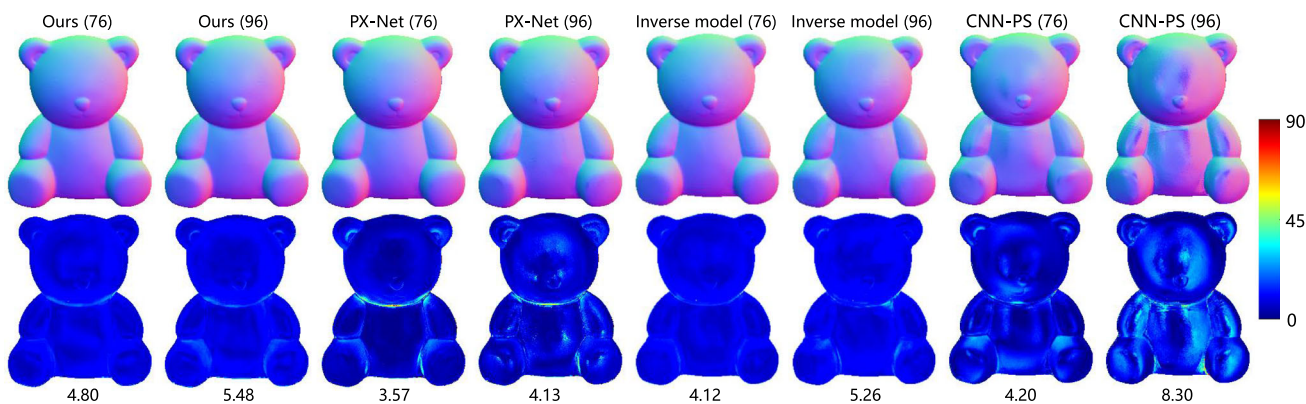


Fig. 11 Results of NormAttention-PSN and CNN-PS (Ikehata, 2018) using 96 input images or 76 input images of the object “Bear”. Numbers below the error maps are the MAE in degrees

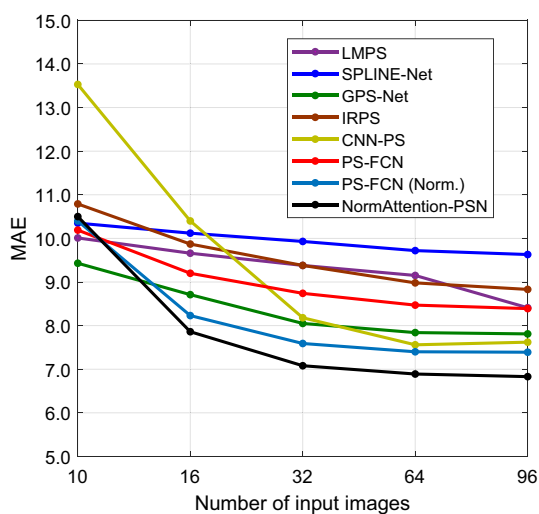


Fig. 12 Comparisons on different numbers of input images

4.4 Evaluation on Different Materials

Fig. 14 shows the results on the object “Dragon”, rendered with 100 different MERL BRDFs (Matusik et al., 2003), and each type of material is tested with 100 images with

random illumination directions in the upper hemisphere. It can be seen that the performance of the estimated surface normals is promising, outperforming the PS-FCN (Norm.) (icitechen2020deep) on most of the materials. Our method achieves an average MAE of 4.65° on all of the 100 kinds of materials. On most kinds of materials, our NormAttention-PSN reconstructs the surface normals with less than 5 degrees angular error, which illustrates the robustness of our method meeting different surface materials.

4.5 Evaluation on the Light Stage Data Gallery

We further evaluated our method on the more complex Light Stage Data Gallery data set (Einarsson et al., 2006), with general non-Lambertian materials. Fig. 15 shows the qualitative results of our method. Similarly, our method was trained with 32 images, while being evaluated with 100 input images randomly selected from 253 images. Note that the input images of the objects “Helmet”, “Plant”, and “Fighting” are down-sampled to half of the spatial resolution, because the original resolution is too large to process.

As shown in Fig. 15, the estimated normals retain the details without blurring, such as the screws of the object “Helmet”, and the lumpy-looking clothes of the objects

Table 6 Comparison of different methods on the DiLiGenT benchmark test data set (Shi et al., 2019). All methods are evaluated with 96 images

Method	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Average
Least square (Woodham, 1980)	9.36	15.15	8.43	21.97	15.86	27.62	10.44	17.06	21.41	16.36
IK12 (Ikehata et al., 2012)	6.81	10.90	7.54	22.17	13.44	26.46	7.44	14.95	18.05	14.20
ST14 (Shi et al., 2014)	6.09	10.92	6.43	10.82	10.33	25.43	6.64	8.97	14.16	11.08
ST12 (Shi et al., 2012)	5.12	11.00	5.61	11.18	10.54	24.82	6.33	8.83	13.27	10.74
DPSN (Santo et al., 2017)	6.32	12.80	5.82	8.00	12.04	17.78	8.26	9.02	16.11	10.68
PS-FCN (Chen et al., 2018)	5.42	8.30	6.24	7.98	8.62	15.93	7.59	7.11	13.43	8.96
PS-FCN (Norm.) (Chen et al., 2020a)	5.40	8.22	4.39	7.44	8.02	12.69	5.83	7.12	11.57	7.85
NormAttention-PSN (ours)	3.77	8.44	4.28	7.67	7.03	12.84	4.96	5.53	11.17	7.30



Fig. 13 Quantitative results on the objects “Cow”, “Goblet”, “Harvest”, and “Reading” in the DiLiGenT test data set (Shi et al., 2019) with 96 input images

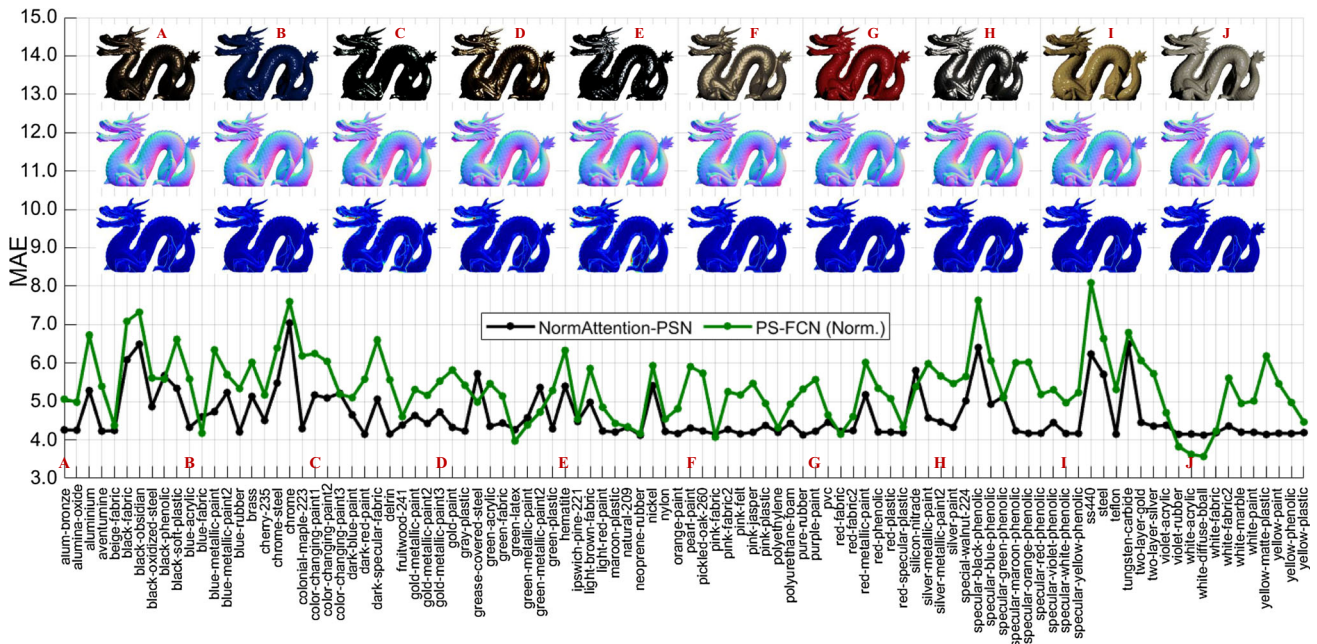


Fig. 14 The MAE of the estimated surface normals on the samples of “Dragon” with 100 kinds of material, from MERL BRDFs (Matusik et al., 2003). We also show the visual results of our method on ten materials, denoted as A ~ J

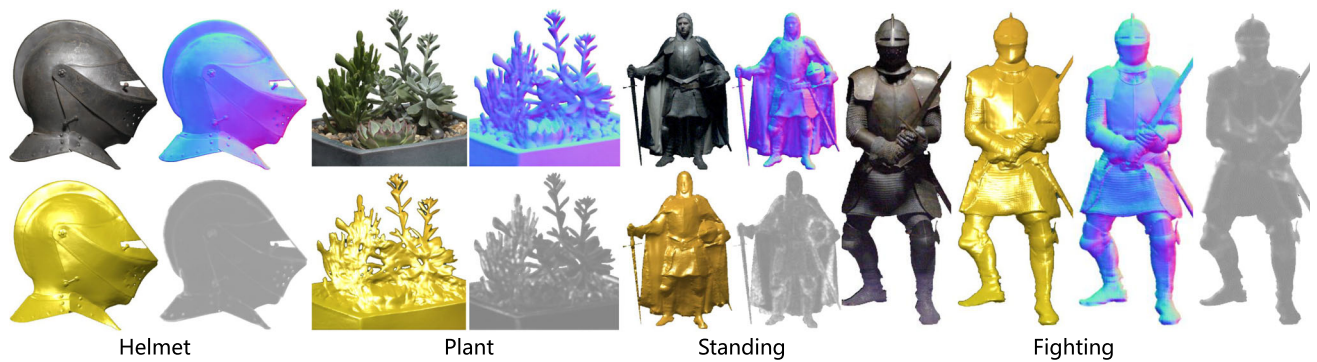


Fig. 15 Qualitative results of our method on four objects, “Helmet”, “Plant”, “Standing”, and “Fighting”, in the Light Stage Data Gallery (Einarsson et al., 2006), with 100 input images

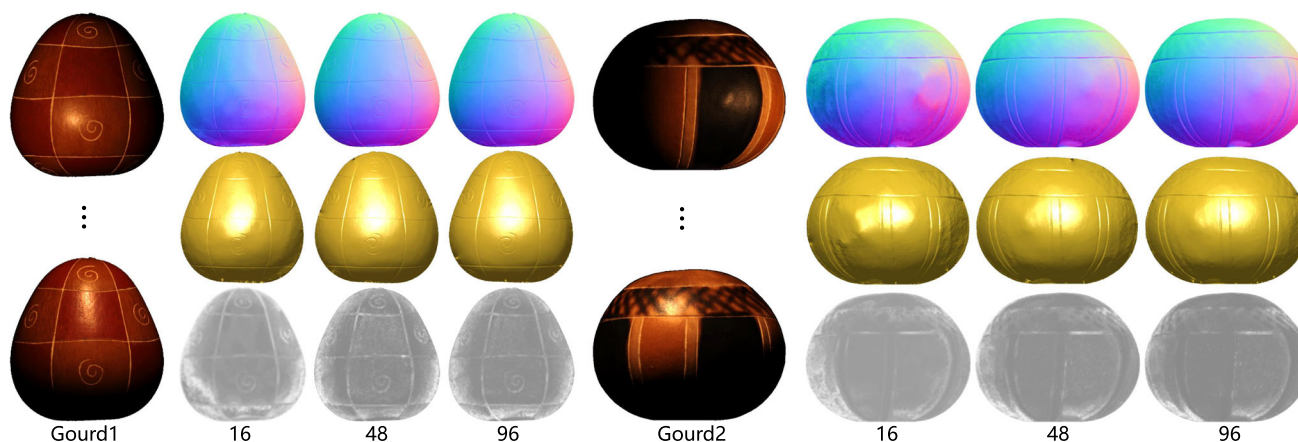


Fig. 16 Qualitative results of our method on the objects, “Gourd1” and “Gourd2”, in the Gourd data set (Alldrin et al., 2008). Next to each of the object observation, from top to bottom, the three rows show the

estimated normal map, the 3D reconstruction, and the attention map, where 16, 48, and 96 represent the numbers of input images

“Standing” and “Fighting”. Furthermore, it can be seen that the attention maps are activated in high-frequency regions, such as edges and crinkles. Note that the BRDFs of the object “Plants” are not used during training. However, the results of “Plant” are quite visually accurate, which shows the robustness of our method.

4.6 Evaluation on the Gourd Data Set

Furthermore, we qualitatively evaluated our method on the Gourd data set (Alldrin et al., 2008). The surfaces of the objects in this data set are associated with crinkles and spatially varying BRDFs. Fig. 16 shows the visual results based on our method, which is trained with 32 images, while tested with 16, 48, and 96 input images.

It can be seen that our method has the flexibility of handling any number of input images. The estimated surface normals for the objects, on “Gourd1” and “Gourd2”, can show clearly crinkles of the objects, under all of the input conditions. The estimation results, based on our method, are robust to spatially varying BRDFs, in particular, on “Gourd2”. Furthermore, the generated attention maps, based on different numbers of input images, can show the correct representation of the complicated structures (crinkles), although some noise can be found in the results when 16 input images are used.

5 Conclusions

In this paper, we proposed a double-gate normalized attention-weighted photometric stereo network, namely NormAttention-PSN, which significantly improves the estimation of surface

normals, especially in those high-frequency regions. We present an attention-weighted loss, which provides an adaptive weight for the detail-preserving gradient loss, to optimize the proposed network. We also employed a double-gate observation normalization strategy to explicitly remove the influence of spatially varying BRDFs and avoid the impact of non-Lambertian surfaces. We further adopted a parallel high-resolution structure to extract features. The ablation experiments have illustrated the effectiveness of the different components of our method, which benefit the estimation of surface normals. Extensive quantitative and qualitative comparisons on both the real (the DiLiGenT benchmark, the Light Stage Data Gallery, and the Gourd data set) and synthetic (the Dragon data set) data sets have shown that our method outperforms previous deep learning-based photometric stereo methods. The visual examples have also demonstrated that our proposed NormAttention-PSN can better predict the surface normals of high-frequency regions, having spatially varying BRDFs, and being non-Lambertian. Furthermore, NormAttention-PSN can provide a framework for other low-level and medium-level regression tasks, such as depth estimation and image enhancement, where the attention-weighted loss can benefit the recovery of structural details.

Acknowledgements The work was supported by the Key-Area Research and Development Program of Guangdong Province (2020B090928001), the Project of Strategic Importance Fund from The Hong Kong Polytechnic University (No. ZE1X), the National Key R&D Program of China under Grant (2018AAA0100602), the National Key Scientific Instrument and Equipment Development Projects of China (41927805), and the National Natural Science Foundation of China (61872012, 62136001, 61976123, 61601427), the Key Development Program for Basic Research of Shandong Province (ZR2020ZD44), and the Taishan Young Scholars Program of Shandong Province.

References

- Ackermann, J., Goesele, M., et al. (2015). A survey of photometric stereo techniques. *Foundations and Trends @ in Computer Graphics and Vision*, 9(3), 149–254.
- Alldrin, N. Zickler, T. & Kriegman, D. (2008). Photometric stereo with non-parametric and spatially-varying reflectance. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE.
- Alldrin, N. G., & Kriegman, D. J. (2007). Toward reconstructing surfaces with arbitrary isotropic reflectance: A stratified photometric stereo approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 1–8). IEEE.
- Barsky, S., & Petrou, M. (2003). The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), 1239–1252.
- Basri, R., & Jacobs, D. W. (2003). Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2), 218–233.
- Blau, Y. & Michaeli, T. (2018). The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6228–6237).
- Chandraker, M. Agarwal, S. & Kriegman, D. (2007). Shadowcuts: Photometric stereo with shadows. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE.
- Chandraker, M., Bai, J., & Ramamoorthi, R. (2012). On differential photometric reconstruction for unknown, isotropic brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2941–2955.
- Chen, G. Han, K. & Wong, K. Y. K. (2018). Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European conference on computer vision* (pp. 3–18).
- Chen, G. Han, K. Shi, B. Matsushita, Y. & Wong, K. Y. K. (2019) Self-calibrating deep photometric stereo networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8739–8747).
- Chen, G., Han, K., Shi, B., Matsushita, Y., & Wong, K. Y. K. (2020). Deep photometric stereo for non-lambertian surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1), 129–142.
- Chen, G. Waechter, M. Shi, B. Wong, K. Y. K. & Matsushita, Y. (2020b). What is learned in deep uncalibrated photometric stereo? In *Proceedings of the European conference on computer vision* (pp. 745–762). Springer.
- Cheng, W. C. (2006). Neural-network-based photometric stereo for 3d surface reconstruction. In *The 2006 IEEE International joint conference on neural network proceedings* (pp. 404–410) IEEE.
- Chung, H. S. & Jia, J. (2008). Efficient photometric stereo on glossy surfaces with wide specular lobes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. (pp. 1–8). IEEE.
- Einarsson, P. Chabert, C. F. Jones, A. Ma, W. C. Lamond, B. Hawkins, T. Bolas, M. Sylwan, S. & Debevec, P. (2006). Relighting human locomotion with flowed reflectance fields. In *proceedings of the eurographics conference on rendering techniques* (pp. 183–194).
- Georgiades, A. S. (2003). Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In: *Proceedings of the IEEE international conference on computer vision* (p. 816).
- Goldman, D. B., Curless, B., Hertzmann, A., & Seitz, S. M. (2010). Shape and spatially-varying brdfs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6), 1060–1071.
- Hartmann, W. Galliani, S. Havlena, M. Van Gool, L. & Schindler, K. (2017). Learned multi-patch similarity. In *Proceedings of the IEEE international conference on computer vision*. (pp. 1586–1594).
- He, K. Zhang, X. Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 770–778).
- Herbort, S. & Wöhler, C. (2011). An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods. *3D Research*, 2(3):4 .
- Higo, T. Matsushita, Y. & Ikeuchi, K. (2010). Consensus photometric stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1157–1164) IEEE
- Holroyd, M., Lawrence, J., Humphreys, G., & Zickler, T. (2008). A photometric approach for estimating normals and tangents. *ACM Transactions on Graphics*, 27(5), 1–9.
- Honzátko, D. Türetken, E. Fua, P. Dunbar, L. A. (2021). Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo. In *Proceedings of the international conference on 3D vision* (pp. 394–402).
- Hui, Z., & Sankaranarayanan, A. C. (2016). Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10), 2060–2073.
- Ikehata, S. (2018). Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proceedings of the European conference on computer vision* (pp. 3–18).
- Ikehata, & S. Aizawa, K. (2014). Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2179–2186).
- Ikehata, S. Wipf, D. Matsushita, Y. & Aizawa, K. (2012). Robust photometric stereo using sparse regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 318–325). IEEE
- Isola, P. Zhu, J. Y. Zhou, T. & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).
- Iwahori, Y. Woodham, R. J. Tanaka, H., & Ishii, N. (1993). Neural network to reconstruct specular surface shape from its three shading images. In *Proceedings of international conference on neural networks 2*, (pp.1181–1184) IEEE.
- Jian, M., Dong, J., Gong, M., Yu, H., Nie, L., Yin, Y., & Lam, K. M. (2019). Learning the traditional art of chinese calligraphy via three-dimensional reconstruction and assessment. *IEEE Transactions on Multimedia*, 22(4), 970–979.
- Johnson, M. K. & Adelson, E. H. (2011). Shape estimation in natural illumination. In *Proceedings of the IEEE international conference on computer vision*, (pp. 2553–2560). IEEE.
- Ju, Y. Jian, M. Dong, J. & Lam, K. M. (2020a). Learning photometric stereo via manifold-based mapping. In: *Proceedings of the IEEE international conference on visual communications and image processing (VCIP)*, (pp. 411–414). IEEE.
- Ju, Y. Lam, K. M. Chen, Y. Qi, L. & Dong, J. (2020b). Pay attention to devils: A photometric stereo network for better details. In: *Proceedings of the twenty-ninth international joint conference on artificial intelligence* (pp. 694–700).
- Ju, Y., Dong, J., & Chen, S. (2021). Recovering surface normal and arbitrary images: A dual regression network for photometric stereo. *IEEE Transactions on Image Processing*, 30, 3676–3690.
- Ju, Y., Peng, Y., Jian, M., Gao, F., & Dong, J. (2022). Learning conditional photometric stereo with high-resolution features. *Computational Visual Media*, 8(1), 105–118.

- Li, J., Robles-Kelly, A., You, S., & Matsushita, Y. (2019). Learning to minify photometric stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7568–7576).
- Logothetis, F., Budvytis, I., Mecca, R., & Cipolla, R. (2021). Px-net: Simple and efficient pixel-wise training of photometric stereo networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 12757–12766).
- Matusik, W., Pfister, H., Brand, M., & McMillan, L. (2003). A data-driven reflectance model. *ACM Transactions on Graphics*, 22(3), 759–769.
- McAuley, S., Hill, S., Hoffman, N., Gotanda, Y., Smits, B., Burley, B., & Martinez, A. (2012). Practical physically-based shading in film and game production. In *ACM SIGGRAPH 2012 Courses* (pp. 1–7).
- Miyazaki, D., Hara, K., & Ikeuchi, K. (2010). Median photometric stereo as applied to the segonko tumulus and museum objects. *International Journal of Computer Vision*, 86(2–3), 229–242.
- Mukaigawa, Y., Ishii, Y., & Shakunaga, T. (2007). Analysis of photometric factors based on photometric linearization. *JOSA A*, 24(10), 3326–3334.
- Nayar, S. K., Ikeuchi, K., & Kanade, T. (1991). Shape from interreflections. *International Journal of Computer Vision*, 6(3), 173–195.
- Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y. (2017). Deep photometric stereo network. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 501–509).
- Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y. (2020). Deep photometric stereo networks for determining surface normal and reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence* p early access.
- Shi, B., Tan, P., Matsushita, Y., Ikeuchi, K. (2012). Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances. In: *Proceedings of the european conference on computer vision* (pp. 455–468). Springer.
- Shi, B., Tan, P., Matsushita, Y., & Ikeuchi, K. (2014). Bi-polynomial modeling of low-frequency reflectances. *IEEE transactions on pattern analysis and machine intelligence*, 36(6), 1078–1091.
- Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S., & Tan, P. (2019). A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 271–284.
- Simchony, T., Chellappa, R., & Shao, M. (1990). Direct analytical methods for solving poisson equations in computer vision problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5), 435–446.
- Solomon, F., & Ikeuchi, K. (1996). Extracting the shape and roughness of specular lobe objects using four light photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4), 449–454.
- Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5693–5703).
- Taniai, T., & Maehara, T. (2018). Neural inverse rendering for general reflectance photometric stereo. In *Proceedings of the international conference on machine learning* (pp. 4857–4866).
- Tozza, S., Mecca, R., Duocastella, M., & Del Bue, A. (2016). Direct differential photometric stereo shape recovery of diffuse and specular surfaces. *Journal of Mathematical Imaging and Vision*, 56(1), 57–76.
- Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T. (2017) Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5038–5047)
- Verbiest, F. & Van Gool, L. (2008). Photometric stereo with coherent outlier handling and confidence estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8).
- Wang, X., Jian, Z., & Ren, M. (2020). Non-lambertian photometric stereo network based on inverse reflectance model with collocated light. *IEEE Transactions on Image Processing*, 29, 6032–6042.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.
- Wiles, O. & Zisserman, A. (2017). Silnet: Single-and multi-view reconstruction by learning from silhouettes. In *Proceedings of the British machine vision conference*.
- Woodham, R. J. (1980). Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1), 139–144.
- Wu, L., Ganesh, A., Shi, B., Matsushita, Y., Wang, Y. & Ma, Y. (2010) Robust photometric stereo via low-rank matrix completion and recovery. In *Proceedings of the asian conference on computer vision* (pp. 703–717). Springer.
- Wu, S., Rupprecht, C. & Vedaldi, A. (2020). Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–10).
- Yao, Z., Li, K., Fu, Y., Hu, H. & Shi, B. (2020). Gps-net: Graph-based photometric stereo network. In *Proceedings of the advances in neural information processing systems*
- Yeung, S. K., Wu, T. P., Tang, C. K., Chan, T. F., & Osher, S. J. (2015). Normal estimation of a transparent object using a video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4), 890–897.
- Yu, C., Seo, Y., Lee, & S. W. (2010). Photometric stereo from maximum feasible lambertian reflections. In: *Proceedings of the European conference on computer vision* (pp. 115–126) Springer.
- Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L. Y. & Kot, A.C. (2019) Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 8549–8558).
- Zheng, Q., Shi, B., & Pan, G. (2020). Summary study of data-driven photometric stereo methods. *Virtual Reality & Intelligent Hardware*, 2(3), 213–221.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.