A summary study of data-driven photometric stereo

Qian ZHENG¹, Boxin SHI²*, Gang PAN³

¹. School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
². National Engineering Laboratory for Video Technology, Department of CS, Peking University, Beijing, China
³. State Key Lab of CAD&CG, College of Computer Science and Technology, Zhejiang University, Hangzhou, China

*Corresponding author, shiboxin@pku.edu.cn

Received: 18 February 2020 Accepted: 18 March 2020

Abstract

Background A photometric stereo method aims to recover the surface normal of a 3D object observed under varying light directions. It is an ill-posed problem due to the unknown of general reflectance properties of the surface.

Methods This paper reviews existing data-driven methods and focuses on their technical insights for the problem of photometric stereo. We divide these methods into two categories according to how they process an image, i.e., in a per-pixel and all-pixel manner. We discuss the differences and relations of methods between the two categories. The discussions are from aspects of inputs, networks, and data, which are key factors of designing a deep learning approach.

Results We demonstrate their performance using a popular benchmark dataset.

Conclusions Data-driven photometric stereo methods have shown superior performance advantage over traditional methods. However, these methods suffer from various limitations, such as limited generalization capability. Finally, this study suggests research directions for future work.

Keywords Photometric Stereo; Data-driven Methods; Non-Lambertian Reflectance

1 Introduction

Modern 3D computer stereo vision methods, including geometric (e.g., binocular¹¹ and multi-view stereo²¹) and photometric approaches³, have achieved faithful 3D reconstruction from a set of images. Photometric methods prevail in producing fine details of the surface at a superior resolution for highly accurate 3D shape reconstruction⁴. Despite its long history in computer vision⁵, photometric stereo (PS) is still a fundamentally challenging research problem due to the unknown reflectance and global illumination effects in real-world objects⁶. Traditional methods address these difficulties by modeling non-Lambertian reflectance using Bidirectional Reflectance Distribution Function (BRDF) (e.g., analytic⁶ or empirical⁷) BRDF representations) and considering global illumination effects as outliers (e.g., Sparse Bayesian
Learning\textsuperscript{[8]}. However, such hand-crafted reflectance models are generally effective for limited categories of reflectance\textsuperscript{[9]}.

Inspired by the powerful modeling capacity of deep neural networks for various computer vision tasks (\textit{e.g.}, light estimation\textsuperscript{[10]}, stereo vision\textsuperscript{[11]}), researchers investigate to learn practical reflectance models through data-driven approaches to solve the problem of photometric stereo. DPSN (Deep Photometric Stereo Network)\textsuperscript{[9]} is the first attempt to address non-Lambertian reflectance using deep learning technologies. This approach enforces the testing probe sharing the same pre-defined set of light directions to that of training data, which limits its generalization, \textit{i.e.}, a new model has to be retrained to test data with different lightings. CNN-PS (Convolutional Neural Network based Photometric Stereo)\textsuperscript{[11]}, PS-FCN (Photometric Stereo using Fully Convolutional Network)\textsuperscript{[12]}, and IRPS (neural Inverse Rendering for general reflectance Photometric Stereo)\textsuperscript{[13]} relax this constraint so that data with order-agnostic light directions can be tested. LMPS (Learning to Minify Photometric Stereo)\textsuperscript{[14]} and SPLINE-Net (Sparse Photometric stereo though Lighting Interpolation and Normal Estimation)\textsuperscript{[15]} further consider a small number of lightings, which help reduce the complexity of the data capture process for this problem. SDPS (Self-calibration Deep Photometric Stereo)\textsuperscript{[16]} assumes uncalibrated lightings and achieves state-of-the-art performance. Moreover, Outdoor-PS (single day Outdoor Photometric Stereo)\textsuperscript{[17]} studies to apply data-driven photometric stereo methods for outdoor scenarios (\textit{i.e.}, a partly cloudy or sunny day).

This paper reviews the recent eight attempts using data-driven methods to solve the problem of photometric stereo, based on our tutorial\textsuperscript{1} and course\textsuperscript{2} in the latest conferences. For a comprehensive discussion of non-learning based photometric stereo methods, we refer readers to survey papers\textsuperscript{[5,18,19]}.

\section*{2 Data-Driven Photometric Stereo Methods}

A recent survey paper\textsuperscript{[5]} divides traditional non-Lambertian photometric stereo methods into outlier rejection based methods\textsuperscript{[8,20]}, analytic BRDF modeling based methods\textsuperscript{[6]}, and empirical BRDF modeling based methods\textsuperscript{[7,21,22]} according to the reflectance model they adopted. These methods can also be categorized as per-pixel\textsuperscript{[6,7,8,21,22]} and all-pixel\textsuperscript{[20]} methods according to how they process the input images, \textit{i.e.}, either observed intensities for a pixel or whole images. We follow this simple strategy and divide data-driven methods into per-pixel methods\textsuperscript{[9,11,14,15]} and all-pixel methods\textsuperscript{[12,13,16,17]}. Figure 1 shows frameworks of a per-pixel method\textsuperscript{[9]} and an all-pixel method\textsuperscript{[12]}. Besides the inputs and networks, Figure 1 also illustrates a training dataset\textsuperscript{[9]} and a testing dataset\textsuperscript{[5]} used by a data-driven method.

Data-driven photometric stereo methods aim to optimize a neural network $f(\cdot)$ such that

\[ N = f(I), \]

where the input $I$ can be either $L$ observed intensities for a specific pixel or $L$ observed images under $L$ light directions\textsuperscript{3}, the output surface normal $N$ is accordingly represented by either a 3-dimensional vector or a

\textsuperscript{1} ICIP 2019 Tutorial. https://www.dropbox.com/s/zbhf2yk71z1q3pei/ICIP2019Tutorial_P3R.pdf?dl=0
\textsuperscript{3} We use $L$ to represent the number of input images.
map with the same resolution of input images, $f(\cdot)$ is optimized through a training dataset. The following discussions try to avoid mathematical notations and focus on insights in aspects of inputs, networks, and data.

### 2.1 Inputs

Per-pixel methods\cite{9,11,14,15} take observed intensities as the input and output a surface normal for a single pixel, while all-pixel methods\cite{12,13,16,17} directly take observed images or patches (multi-pixels) as the input and output a surface normal map with the same resolution of its input. This difference indicates that per-pixel methods aim to fit an accurate reflectance model for each pixel while all-pixel methods focus on extracting accurate surface normal maps from various appearances.

**Per-pixel observation as intensity profile.** Intensity profile is a $L$-dimension vector that orders elements in observed intensities by the indices of light directions. As the information of light directions is not fully used during network training, methods with this input \textit{(i.e., DPSN)} assumes light directions between testing and training data are the same. This strong assumption limits the generalization, \textit{i.e.,} a new model has to be retrained to test data with different lightings.

**Per-pixel observation as an observation map.** The observation map\cite{11} is proposed to overcome the above shortcoming (Figure 2). The observation map rearranges observation intensities according to light directions, \textit{i.e.,} directly encodes light directions as 2D coordinates and accordingly projects observation intensities to a 2D space. CNN-PS\cite{11}, LMPS\cite{14}, and SPLINE-Net\cite{15} adopt this data structure to obtain inputs of their neural networks. As the information of light directions is fully kept, CNN-PS\cite{11}, LMPS\cite{14}, and SPLINE-Net\cite{15} can handle inputs with order-agnostic lightings.
All-pixel observation using the whole image. The input of Outdoor-PS\textsuperscript{[17]} consists of \( L \) 16 × 16 image patches. These patches are ordered according to timestamps in a day (different light directions can be observed in different timestamps). Therefore, the light directions are not fully used during network training and Outdoor-PS\textsuperscript{[17]} suffers from a similar limitation as that of DPSN\textsuperscript{[9]} i.e., a new model has to be retrained to test data with different lightings.

All-pixel observation using patches. Other all-pixel methods take \( L \) whole images\textsuperscript{[16]} (or as well as its corresponding light direction\textsuperscript{[12,13]} as the input. To test data with order-agnostic lightings, these methods either leverage a sharing weights scheme\textsuperscript{[12,16]} or impose an unsupervised manner\textsuperscript{[13]}. Besides the classical photometric stereo, several recent advances in multispectral photometric stereo are also with the all-pixel manner, e.g., Antensteiner et al.\textsuperscript{[32]} take the whole image as the input and estimate surface normal through a U-Net\textsuperscript{[34]}, Ju et al.\textsuperscript{[33]} first adopt image patches to estimate a coarse surface normal map and then refine it by a per-pixel pixel manner. Because multispectral photometric stereo involves two additional challenges as compared with the classical one\textsuperscript{[3]}; 1) ambiguity brought by the decomposition of spectra and 2) an extremely small number of observations (i.e., often 3), discussions in the following focuses on photometric stereo without relying on spectral constraints.

2.2 Networks

We brief the overall architectures and specific designs for these data-driven photometric stereo methods in this section. We refer readers to relevant papers\textsuperscript{[9,11,12,13,14,15,16,17]} for more details.

Overall architectures. Expect for DPSN\textsuperscript{[9]} that leverages a classical deep neural network (DNN) architecture (consists of an input layer, hidden layers, and an output layer), other data-driven methods impose the architecture of convolutional neural network (CNN). This is because both observation maps and natural images share the property of spatial continuity (Figure 2). For per-pixel methods\textsuperscript{[9,11,14,15]}, variations of DenseNet\textsuperscript{[23]} are utilized to generate a 3-dimensional surface normal. Because DenseNet\textsuperscript{[23]} are expected to strengthen feature propagation and encourage feature reuse for low dimension (i.e., 3) outputs\textsuperscript{[23]}. For all-pixel methods\textsuperscript{[12,13,16,17]}, the idea of sharing weights among different modules are used. Because this design benefits to aggregating features extracted from multi-observations (e.g., PS-FCN\textsuperscript{[12]} in Figure 1) or enriching the feature extracted from a single observation\textsuperscript{[13,16]}. Architectures of easier setups (i.e., known and large number of lightings\textsuperscript{[9,11,12,17]} are less complicated, while those of difficult ones (i.e., PS with a small number of light directions\textsuperscript{[15]}, unknown lightings\textsuperscript{[16]}, or in an unsupervised manner\textsuperscript{[13]}) contain two sub-networks for joint optimization (e.g., SDPS\textsuperscript{[16]} in Figure 3).
The framework of SDPS\cite{16} which consists of a lighting calibration network and a normal estimation network.

Figure 3 Illustration of symmetric (①②) and asymmetric patterns (③④⑤⑥) on observation maps (from\cite{15}). Red lines represent the symmetric axis.

**Specific designs.** Despite cues that implicitly learned from training data, data-driven methods design specific modules or loss functions to improve the robustness of photometric stereo. Per-pixel methods explicitly leverage properties of the general reflectance (*i.e.*, isotropic BRDF and global illumination effects), while all-pixel methods focus on effectively regressing appearances to shapes in an end-to-end scheme.

- **Isotropic BRDF.** The consideration of isotropy benefits to narrow down the solution space during the training of neural networks. CNN-PS\cite{11} shows that the observation map is rotational pseudo-invariance to the surface normal, based on which additional data are augmented for training (by rotating the observation map and its surface normal simultaneously). SPLINE-Net\cite{15} further discovers that an ideal observation map (*i.e.*, without global illumination effects) is with a symmetric pattern, based on which a symmetric loss function is proposed (Figure 4).

- **Global illumination effects.** Taking global illumination effects such as cast shadow and inter-reflection into account helps for robust estimation for data in the real world. To simulate cast shadow during training, DPSN\cite{9} adopts the operation of dropout, LMPS\cite{14} annotates the observation map by enforcing a part of intensities to be zero. SPLINE-Net\cite{15} shows that the global illumination effects can break the symmetry of an observation map (Figure 4), based on which they introduce an asymmetric loss function.

- **End-to-end.** All-pixel methods pay more attention to information transferring among different modules in a high level view. These methods concatenate all observed images\cite{13,16,17} and/or aggregated all feature...
(from all images)\cite{12,16} to extract shape information, as all images contain the same surface normal map. Some of these methods further extract reflectance\cite{13} (or lighting\cite{16}) information from a single image and shape information, as reflectance (or lighting) information is jointly determined by the single observation and the shape.

2.3 Data

As the data of outdoor PS methods\cite{17} are quite different from those of indoor ones, this section focuses on training and testing datasets used in indoor PS methods.

Training datasets. It is difficult to capture large scale data with the ground truth of surface normal. Therefore, most\cite{4} of data-driven photometric stereo methods synthesize data for training. Table 1 details shapes, materials, light configurations, and rendering engines when synthesizing training data for different methods. As can be observed, 3D models are from datasets of Blobby Shape\cite{24}, Sculpture Shape\cite{25} or the internet. Surface materials are approximated using MERL BRDF\cite{26} or Disney’s principled BSDF\cite{27}. Light directions are set to be fixed (the same as those of testing data), uniformly sampled, or randomly sampled. Engines of Mitsuba\cite{28} or Cycles\cite{29} (e.g., Outdoor-PS\cite{17}) are employed for rendering. Table 1 also displays the numbers of shapes, materials, light directions and images used by each method. It can be found that per-pixel methods\cite{9,11,14,15} require a small number of shape models (i.e., smaller than 15) while all-pixel ones\cite{12,16} impose a much larger number of shape models (i.e., about 42K). This observation is consistent with our discussion in Section 2.1 that per-pixel methods aim to fit an accurate reflectance model for each pixel while all-pixel methods focus on extracting accurate surface normal maps from various appearances.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Details of training data regarding shape, material, light, images for different data-driven photometric stereo methods.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Shape (Number)</td>
</tr>
<tr>
<td>DPSN\cite{9}</td>
<td>Blobby\cite{24} (8)</td>
</tr>
<tr>
<td>CNN-PS\cite{11}, SPLINE-Net\cite{15}</td>
<td>MERL\cite{29} (100)</td>
</tr>
<tr>
<td>LMPS\cite{9}</td>
<td>Fixed (96)</td>
</tr>
<tr>
<td>PS-FCN\cite{12}, SDPS\cite{14}</td>
<td>Mitsuba\cite{29} (76800)</td>
</tr>
<tr>
<td></td>
<td>Material (Number)</td>
</tr>
<tr>
<td></td>
<td>Internet (15)</td>
</tr>
<tr>
<td></td>
<td>Disney\cite{27} (15000)</td>
</tr>
<tr>
<td></td>
<td>Uniform (1300)</td>
</tr>
<tr>
<td></td>
<td>Cycles\cite{29} (19500)</td>
</tr>
<tr>
<td></td>
<td>Light (Number)</td>
</tr>
<tr>
<td></td>
<td>Bloppy\cite{24} (9)</td>
</tr>
<tr>
<td></td>
<td>MERL\cite{29} (100)</td>
</tr>
<tr>
<td></td>
<td>Random (144)</td>
</tr>
<tr>
<td></td>
<td>Mitsuba\cite{29} (10368)</td>
</tr>
<tr>
<td></td>
<td>Image (Number)</td>
</tr>
<tr>
<td></td>
<td>Blobby\cite{24} &amp; Sculpture\cite{29} (~42K)</td>
</tr>
<tr>
<td></td>
<td>MERL\cite{15} (100)</td>
</tr>
<tr>
<td></td>
<td>Random (64)</td>
</tr>
<tr>
<td></td>
<td>Mitsuba\cite{29} (~5.4M)</td>
</tr>
</tbody>
</table>

Testing datasets. The DiLiGenT dataset\cite{5} is the most widely used real-world dataset for evaluation. It consists of 10 different objects with different scales of non-Lambertian reflectance (Figure 1). Each object is illuminated and photographed under 96 different lighting directions. The ground truth of surface normal maps is also provided. Table 2 displays the quantitative results from these data-driven methods as well as a traditional photometric stereo method ST14\cite{22} that achieves state-of-the-art performance. Note that inputs of these methods are quite different, i.e., DPSN\cite{9}, CNN-PS\cite{11}, PS-FCN\cite{12}, IRPS\cite{13}, ST14\cite{22} take 96 images with known light directions as inputs, SPLINE-Net\cite{15} takes 10 random ones, LMPS\cite{14} takes 10 optimal ones, and SDPS\cite{16} takes 96 images without lighting information. As can be observed, results from deep learning based methods (i.e., DPSN\cite{9}, CNN-PS\cite{11}, PS-FCN\cite{12}, IRPS\cite{13}) achieve much better performance as

\footnote{Note that IRPS\cite{13} does not require any synthetic training data because it is an unsupervised method.}
traditional methods with the same inputs. Moreover, deep learning based methods achieve comparable or better results as compared with traditional methods with a more difficult setting (i.e., a small number of inputs\textsuperscript{[14,15]}, or uncalibrated lightings\textsuperscript{[16]})\textsuperscript{,5}

Besides the quantitative evaluation, the visual quality is also conducted on real data such as GOUDRD&APPLE\textsuperscript{[30]} and Light Stage Data Gallery\textsuperscript{[31]}. In addition, there are several synthetic datasets for validation\textsuperscript{[9,11,12,15,16,17]}. These synthetic data are generally rendered using the same way as their training data.

### 3 Discussion

Despite the state-of-the-art performance achieved by data-driven methods, they suffer from limitations such as expensive computation for testing\textsuperscript{[13]}, sensitivity to global illumination effects\textsuperscript{[11,15]}, constrained light directions\textsuperscript{[9,14,17]}, and uniform materials of surface\textsuperscript{[12,16,17]}. A brief summary is made in Table 3. As can be observed, per-pixel methods pay more attention to the modeling of general BRDF and they are robust to non-uniform distributions of surface materials, however, they perform less optimistic for regions with global illumination effects as the shape information is not explicitly considered. All-pixel methods are commonly trained using various shapes with a uniform material for each shape. Therefore, they produce accurate results

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Architecture</th>
<th>General BRDF</th>
<th>End-to-end</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPSN\textsuperscript{[9]}</td>
<td>Per-pixel</td>
<td>Intensities</td>
<td>CNN</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>CNN-PS\textsuperscript{[11]}</td>
<td>Per-pixel</td>
<td>Intensities</td>
<td>CNN</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>SPLINE-Ne\textsuperscript{[15]}</td>
<td>Per-pixel</td>
<td>Intensities</td>
<td>CNN</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>LMPS\textsuperscript{[14]}</td>
<td>Per-pixel</td>
<td>Intensities</td>
<td>CNN</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>IRPS\textsuperscript{[13]}</td>
<td>All-pixel</td>
<td>Images</td>
<td>CNN</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PS-FCN\textsuperscript{[12]}</td>
<td>All-pixel</td>
<td>Images</td>
<td>CNN</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SDPS\textsuperscript{[14]}</td>
<td>All-pixel</td>
<td>Images</td>
<td>CNN</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Outdoor-PS\textsuperscript{[17]}</td>
<td>All-pixel</td>
<td>Images</td>
<td>CNN</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Method & Input & Architecture & General BRDF & End-to-end & Limitation &
\hline
DPSN\textsuperscript{[9]} & Per-pixel & Intensities & CNN & ✓ & × & Constrained lightings &
\hline
CNN-PS\textsuperscript{[11]} & Per-pixel & Intensities & CNN & ✓ & × & Sensitivity to global illumination effects &
\hline
SPLINE-Ne\textsuperscript{[15]} & Per-pixel & Intensities & CNN & ✓ & × & Sensitivity to global illumination effects &
\hline
LMPS\textsuperscript{[14]} & Per-pixel & Intensities & CNN & ✓ & × & Constrained lightings &
\hline
IRPS\textsuperscript{[13]} & All-pixel & Images & CNN & ✓ & ✓ & Expensive computation &
\hline
PS-FCN\textsuperscript{[12]} & All-pixel & Images & CNN & ✓ & ✓ & Uniform materials &
\hline
SDPS\textsuperscript{[14]} & All-pixel & Images & CNN & ✓ & ✓ & Uniform materials &
\hline
Outdoor-PS\textsuperscript{[17]} & All-pixel & Images & CNN & ✓ & ✓ & Constrained lightings & uniform materials &
\hline
\end{tabular}

\textsuperscript{5} Please refer https://sites.google.com/site/photometricstereodata/single?authuser=0 for more results of traditional methods.
for regions with shadows or inter-reflection, while they are less powerful for objects with non-uniform materials. Based on the discussion above, we suggest future research directions as follows:

- **Combination.** This study demonstrates the unique characteristics of per-pixel methods and all-pixel ones. Mutually combining these two types of approaches to further improve performance will be another exploration.

- **Lambertian reflectance.** A recent work\cite{15} suggests that deep learning based methods generally produce unsatisfied results for Lambertian reflectance which is not comparable with the baseline method\cite{3}. Therefore, reflectance less similar to DiLiGenT\cite{5} or MERL\cite{26} should be considered in a data-driven method to avoid overfitting.

- **Practicality.** As data-driven methods have achieved promising performance in a lab environment, leveraging deep learning techniques to solve the photometric stereo problem in a more under-constrained scenario (e.g., Outdoor-PS\cite{17}) should be considered.

---

**References**


DOI:10.1109/CVPR.2019.00707

DOI: 10.1007/978-3-030-01267-0_1

DOI:10.1007/978-3-030-01240-3_1


DOI:10.1109/CVPR.2019.00775


DOI:10.1109/CVPR.2019.00894

DOI:10.1109/TPAMI.2019.2962693

DOI:10.1561/0600000065

19 Herbort S, Wöhler C. An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods. 3D Research, 2011, 2(3):4
DOI:10.1007/3DRes.03(2011)4

DOI:10.1007/978-3-642-19318-7_55

DOI:10.1109/TIP.2019.2894963

DOI:10.1109/TPAMI.2013.196

DOI:10.1109/CVPR.2017.243

DOI:10.1109/CVPR.2011.5995510
DOI:10.5244/C.31.99
DOI:10.1145/1201775.882343
DOI:10.1109/CVPR.2008.4587656
DOI:10.1016/j.patcog.2019.107162
DOI:10.1007/978-3-319-24574-4_28