Neural Inverse Rendering for General Reflectance Photometric Stereo

Tatsunori Taniai [1] **Takanori Maehara** [1]

Abstract

We present a novel convolutional neural network architecture for photometric stereo (Woodham, 1980), a problem of recovering 3D object surface normals from multiple images observed under varying illuminations. Despite its long history in computer vision, the problem still shows fundamental challenges for surfaces with unknown general reflectance properties (BRDFs). Leveraging deep neural networks to learn complicated reflectance models is promising, but studies in this direction are very limited due to difficulties in acquiring accurate ground truth for training and also in designing networks invariant to permutation of input images. In order to address these challenges, we propose a physics based unsupervised learning framework where surface normals and BRDFs are predicted by the network and fed into the rendering equation to synthesize observed images. The network weights are optimized during testing by minimizing reconstruction loss between observed and synthesized images. Thus, our learning process does not require ground truth normals or even pre-training on external images. Our method is shown to achieve the state-of-the-art performance on a challenging real-world scene benchmark.

1. Introduction

3D shape recovery from images is a central problem in computer vision. While geometric approaches such as binocular (Kendall et al., 2017; Taniai et al., 2017) and multi-view stereo (Furukawa & Ponce, 2010) use images from different viewpoints to triangulate 3D points, photometric stereo (Woodham, 1980) uses varying shading cues of multiple images to recover 3D surface normals. It is well known that photometric methods prevail in recovering fine details of surfaces, and play an essential role for highly accurate 3D shape recovery in combined approaches (Nehab et al., 2005; Esteban et al., 2008; Park et al., 2017). Although there exists a closed-form least squares solution to the simplest Lambertian surfaces, such ideally diffuse materials rarely exist in the real word. Photometric stereo for surfaces with unknown general reflectance properties (*i.e.*, bidirectional reflectance distribution functions or BRDFs) still remains as a fundamental challenge (Shi et al., 2018).

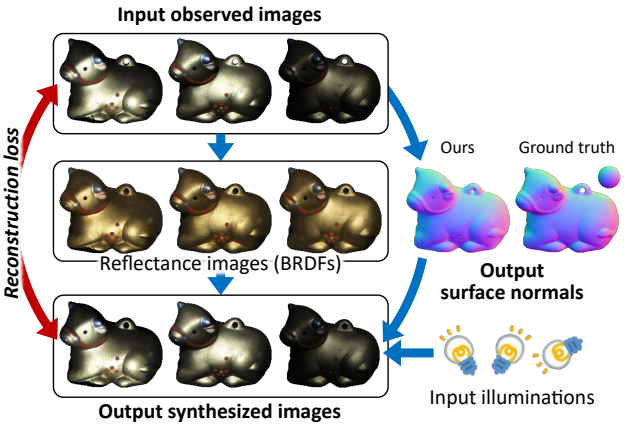


Figure 1. **Reconstruction based photometric stereo.** Given multiple images observed under varying illuminations, our inverse-rendering network estimates a surface normal map and reflectance images. We then reconstruct (or render) the observed images using these estimates and input illuminations. The synthesized images are used to define reconstruction loss for unsupervised learning.

Meanwhile, deep learning technologies have drastically pushed the envelope of state-of-the-art in many computer vision tasks such as image recognition (Krizhevsky et al., 2012; He et al., 2015; 2016), segmentation (He et al., 2017b) and stereo vision (Kendall et al., 2017). As for photometric stereo, it is promising to replace hand-crafted reflectance models with deep neural networks to learn complicated BRDFs. However, studies in this direction so far are surprisingly limited (Santo et al., 2017; Hold-Geoffroy et al., 2018). This is possibly due to difficulties of making a large amount of training data with ground truth. Accurately measuring surface normals of real objects is very difficult, because we need highly accurate 3D shapes to reliably compute surface gradients. In fact, a real-world scene benchmark

[1]RIKEN Center for Advanced Intelligence Project (RIKEN AIP), Nihonbashi, Tokyo, Japan. Correspondence to: Tatsunori Taniai <tatsunori.taniai@riken.jp>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

of photometric stereo with ground truth has only recently been introduced by precisely registering laser-scanned 3D meshes onto 2D images (Shi et al., 2018). Using synthetic training data is possible (Santo et al., 2017), but we need photo-realistic rendering that should ideally account for various realistic BRDFs and object shapes, spatially-varying BRDFs and materials, presence of cast shadows and interreflections, etc. This is more demanding than training-data synthesis for stereo and optical flow (Mayer et al., 2016) where rendering by the simplest Lambertian reflectance often suffices. Also, measuring BRDFs of real materials requires efforts and an existing BRDF database (Matusik et al., 2003) provides only a limited number of materials.

As another difficulty of applying deep learning to photometric stereo, when networks are pre-trained, they need to be invariant to permutation of inputs, *i.e.*, permuting input images (and corresponding illuminations) should not change the resulting surface normals. Existing neural network methods (Santo et al., 2017) avoid this problem by assuming the same illumination patterns throughout training and testing phases, which limits application scenarios of methods.

In this paper, we propose a novel convolutional neural network (CNN) architecture for general BRDF photometric stereo. Given observed images and corresponding lighting directions, our network *inverse renders* surface normals and spatially-varying BRDFs from the images, which are further fed into the reflectance (or rendering) equation to synthesize observed images (see Fig. 1). The network weights are optimized by minimizing reconstruction loss between observed and synthesized images, enabling unsupervised learning that does not use ground truth normals. Furthermore, learning is performed directly on a test scene during the testing phase without any pre-training. Therefore, the permutation invariance problem does not matter in our framework. Our method is evaluated on a challenging real-world scene benchmark (Shi et al., 2018) and is shown to outperform state-of-the-art learning-based (Santo et al., 2017) and other classical unsupervised methods (Shi et al., 2014; 2012; Ikehata & Aizawa, 2014; Ikehata et al., 2012; Wu et al., 2010; Goldman et al., 2010; Higo et al., 2010; Alldrin et al., 2008). We summarize the advantages of our method as follows.

- Existing neural network methods require pre-training using synthetic data, whenever illumination conditions of test scenes change from the trained ones. In contrast, our physics-based approach can directly fit network weights for a test scene in an unsupervised fashion.
- Compared to classical physics-based approaches, we leverage deep neural networks to learn complicated reflectance models, rather than manually analyzing and inventing reflectance properties and models.
- Yet, our physics-based network architecture allows us to exploit prior knowledge about reflectance properties that have been broadly studied in the literature.

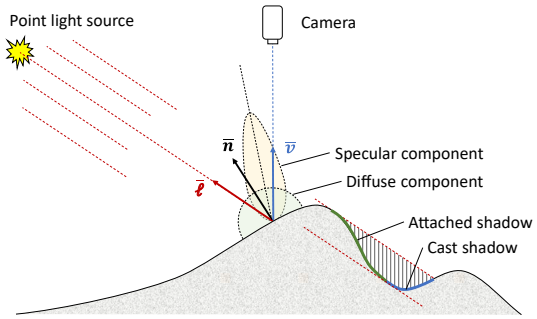


Figure 2. **Surface reflectance and BRDFs.** We illustrate a situation where an object surface point with a normal vector $\bar{\boldsymbol{n}}$ is illuminated by an infinitely distant point light source in a direction $\bar{\boldsymbol{\ell}}$, and is observed by a camera in a view direction $\bar{\boldsymbol{v}}$. Unknown BRDFs have major components of diffuse and specular reflections. Shadows occur at surfaces where $\bar{\boldsymbol{\ell}}^T\bar{\boldsymbol{n}} \leq 0$ (attached shadow) or the light ray is occluded by objects (cast shadow).

2. Preliminaries

Before presenting our method, we recap basic settings and approaches in photometric stereo. Suppose a reflective surface with a unit normal vector $\bar{\boldsymbol{n}} \in \mathbb{R}^3$ is illuminated by a point light source $\boldsymbol{\ell} \in \mathbb{R}^3$ (where $\boldsymbol{\ell} = \ell \cdot \bar{\boldsymbol{\ell}}$ has an intensity $\ell > 0$ and a unit direction $\bar{\boldsymbol{\ell}}$), without interreflection and ambient lighting. When this surface is observed by a linear-response camera in a view direction $\bar{\boldsymbol{v}} \in \mathbb{R}^3$, its pixel intensity $I \in \mathbb{R}_+$ is determined as follows.

$$I = s\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}}) \max(\boldsymbol{\ell}^T\bar{\boldsymbol{n}}, 0) \tag{1}$$

Here, $s \in \{0, 1\}$ is a binary function for the presence of a cast shadow, $\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}})$ is a BRDF, and $\max(\cdot, 0)$ represents an attached shadow. Figure 2 illustrates this situation.

The goal of photometric stereo is to recover the surface normal $\bar{\boldsymbol{n}}$ from intensities I, when changing illuminations $\boldsymbol{\ell}$. Here, we usually assume a camera with a fixed viewpoint and an orthogonal projection model, in which case the view direction $\bar{\boldsymbol{v}}$ is constant and typically $\bar{\boldsymbol{v}} = (0, 0, 1)^T$. Also, light sources are assume to be infinitely distant so that $\boldsymbol{\ell}$ is uniform over the entire object surfaces.

2.1. Lambertian model and least squares method

When the BRDF $\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}})$ is constant as ρ_0, the surface is purely diffuse. Such a model is called the Lambertian reflectance and the value ρ_0 is called albedo. In this case, estimation of $\bar{\boldsymbol{n}}$ is relatively easy, because for bright pixels ($I > 0$) the reflectance equation of Eq. (1) becomes a linear equation: $I = \boldsymbol{\ell}^T\boldsymbol{n}$ where $\boldsymbol{n} = \rho_0\bar{\boldsymbol{n}}$. Therefore, if we know at least three intensity measurements $\boldsymbol{I} \in \mathbb{R}_+^M$ ($M \geq 3$) and their lighting conditions $\boldsymbol{L} = [\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_M]^T \in \mathbb{R}^{M\times 3}$, then we obtain a linear system

$$\boldsymbol{I} = \boldsymbol{L}\boldsymbol{n}, \tag{2}$$

which is solved by least squares as

$$\boldsymbol{n} = \boldsymbol{L}^{\dagger}\boldsymbol{I}. \tag{3}$$

Here, $\boldsymbol{L}^{\dagger}$ is the pseudo inverse of $\boldsymbol{L}$, and the resulting vector $\boldsymbol{n} = \rho_0\bar{\boldsymbol{n}}$ is then L2-normalized to obtain the unit normal $\bar{\boldsymbol{n}}$.

In practice, images are contaminated as $\boldsymbol{I} + \boldsymbol{e}$ due to sensor noises, interreflections, etc. Therefore, we often set a threshold τ for selecting inlier observation pixels $I_i > \tau$.

When the lighting conditions $\boldsymbol{L}$ are unknown, the problem is called *uncalibrated photometric stereo*. It is known that the problem has the so-called *bas-relief ambiguity* (Belhumeur et al., 1999), and is difficult even for the Lambertian surfaces. In this paper, we focus on the *calibrated photometric stereo* settings that assume known lighting conditions.

2.2. Photometric stereo for general BRDF surfaces

When the BRDF $\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}})$ has unknown non-Lambertian properties, photometric stereo becomes very challenging, because we essentially need to know the form of the BRDF $\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}})$ by assuming some reflectance model to it or by directly estimating $\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}})$ along with the surface normal $\bar{\boldsymbol{n}}$. Below we briefly review existing such approaches and their limitations. For more comprehensive reviews, please refer to a recent excellent survey by Shi et al. (2018).

Ourlier rejection based methods. A group of methods treat non-Lambertian reflectance components including specular highlights and shadows as outliers to the Lambertian model. Thus, Eq. (2) is rewritten to

$$\boldsymbol{I} = \boldsymbol{L}\boldsymbol{n} + \boldsymbol{e}, \tag{4}$$

where non-Gaussian outliers $\boldsymbol{e}$ are assume to be sparse. Recent methods solve this sparse regression problem by using robust statistical techniques (Wu et al., 2010; Ikehata et al., 2012) or using learnable optimization networks (Xin et al., 2016; He et al., 2017a). However, this approach cannot handle broad and soft specularity due to the collapse of the sparse outlier assumption (Shi et al., 2018).

Analytic BRDF models. Another type of methods use more realistic BRDF models than the Lambertian model matured in the computer graphics literature, *e.g.*, the Torrance-Sparrow model (Georghiades, 2003), the Ward model (Chung & Jia, 2008), or a Ward mixture model (Goldman et al., 2010). These models explicitly consider specularity rather than treating it as outliers, and often take a form of the sum of diffuse and specular components as follows.

$$\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}}) = \rho_{\text{diff}} + \rho_{\text{spec}}(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}}) \tag{5}$$

However, these methods rely on hand-crafted models that can only handle narrow classes of materials.

General isotropic BRDF properties. More advanced methods directly estimate the unknown BRDF $\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}})$ by exploiting some general BRDF properties. For example, many materials have an isotopic BRDF that only depends on relative angles between $\bar{\boldsymbol{n}}$, $\bar{\boldsymbol{\ell}}$ and $\bar{\boldsymbol{v}}$. Given the isotropy, Ikehata & Aizawa (2014) further assume the following bivariate BRDF function

$$\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}}) = \rho(\bar{\boldsymbol{n}}^T\bar{\boldsymbol{\ell}}, \bar{\boldsymbol{\ell}}^T\bar{\boldsymbol{v}}) \tag{6}$$

with monotonicity and non-negativity constraints. Similarly, Shi et al. (2014) exploit a low-frequency prior of BRDFs and propose a bi-polynomial BRDF:

$$\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}}) = \sum_{i=0}^{k}\sum_{j=0}^{k} C_{ij}x^i y^j, \tag{7}$$

where $x = \bar{\boldsymbol{n}}^T\bar{\boldsymbol{h}}$, $y = \bar{\boldsymbol{\ell}}^T\bar{\boldsymbol{h}}$, and $\bar{\boldsymbol{h}} = (\bar{\boldsymbol{\ell}} + \bar{\boldsymbol{v}})/\|\bar{\boldsymbol{\ell}} + \bar{\boldsymbol{v}}\|_2$.

Our method is close to the last approach in that we learn broad classes of a BRDF from observations without restricting it to a particular reflectance model. However, unlike those methods that fully rely on careful human analysis of BRDF properties, we leverage the powerful expressibility of deep neural networks to learn general complicated BRDFs. Yet, our network architecture also explicitly uses the physical reflectance equation of Eq. (1) internally, which allows us to incorporate abundant wisdom about reflectance developed in the literature, into neural network based approaches.

3. Proposed method

In this section, we present our novel inverse-rendering based neural network architecture for photometric stereo, and explain its learning procedures with a technique of early-stage weak supervision. Here, as standard settings of calibrated photometric stereo, we assume M patterns of light source directions $\boldsymbol{\ell}_i$ ($i \in \{1, 2, \ldots, M\}$) and corresponding image observations $\boldsymbol{I}_i$ as inputs. We also assume that the mask $\boldsymbol{O}$ of target object regions is provided. Our goal is to estimate the surface normal map $\bar{\boldsymbol{N}}$ of the target object regions.

Notations. We use bold capital letters for tensors and matrices, and bold small letters for vectors. We use tensors of dimensionality $D \times H \times W$ to represent images, and normal and other feature maps, where D is some channel number and $H \times W$ is the spatial resolution. Thus, $\boldsymbol{I}_i \in \mathbb{R}^{C\times H\times W}$ and $\bar{\boldsymbol{N}} \in \mathbb{R}^{3\times H\times W}$, where C is the number of color channels of images. We use the subscript p to denote a pixel location of such tensors, *e.g.*, $\bar{\boldsymbol{N}}_p \in \mathbb{R}^3$ is a normal vector at p. The light vectors $\boldsymbol{\ell}_i$ can also have color channels, in which case $\boldsymbol{\ell}_i \in \mathbb{R}^{3\times C}$ are matrices but we use a small letter for intuitiveness. The index i is always used to denote the observation index $i \in \{1, 2, \ldots, M\}$. When we use tensors of dimensionality $B \times D \times H \times W$, the first dimension B denotes a minibatch size processed in one SGD iteration.

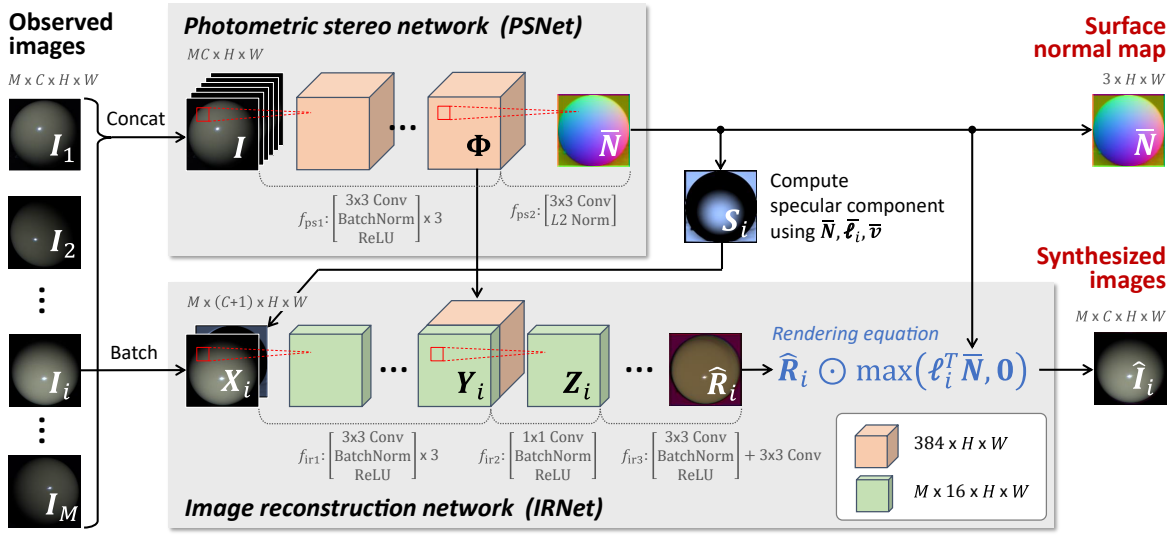


Figure 3. **Proposed network architecture for photometric stereo.** We use two subnetworks, both are fully convolutional. [TOP] The photometric stereo network (PSNet) outputs a surface normal map $\bar{\boldsymbol{N}}$ as the desired solution, given an image tensor $\boldsymbol{I}$ that concatenates all observed images $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_M\}$ of a test scene. [BOTTOM] The image reconstruction network (IRNet) synthesizes each observed image $\boldsymbol{I}_i$ using the rendering equation. IRNet is used to define reconstruction loss between the observed and synthesized images for unsupervised learning. Note that, as calibrated photometric stereo, the lighting directions $\{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_M\}$ are also provided as inputs, and used for the computations of the rendering equation and a specular component input $\boldsymbol{S}_i$. Also, dimensionality $B \times D \times H \times W$ of tensors denotes a minibatch size B, channel number D, and spatial resolution $H \times W$, respectively, where $B = 1$ is omitted in PSNet.

3.1. Network architecture

We illustrate our network architecture in Fig. 3. Our method uses two subnetworks, which we name the photometric stereo network (PSNet) and image reconstruction network (IRNet), respectively. PSNet predicts a surface normal map as the desired output, given the input images. On the other hand, IRNet synthesizes observed images using the rendering equation of Eq. (1). The synthesized images are used to define reconstruction loss with the observed images, which produces gradients flowing into both networks and enables learning without ground truth supervision. We now explain these two networks in more details below.

3.1.1. PHOTOMETRIC STEREO NETWORK

Given a tensor $\boldsymbol{I} \in \mathbb{R}^{MC \times H \times W}$ that concatenates all M input images along the channel axis, PSNet first converts it to an abstract feature map $\boldsymbol{\Phi} \in \mathbb{R}^{D_{\text{ps}} \times H \times W}$ as

$$\boldsymbol{\Phi} = f_{\text{ps1}}(\boldsymbol{I}; \boldsymbol{\theta}_{\text{ps1}}), \tag{8}$$

and then outputs a surface normal map $\bar{\boldsymbol{N}}$ given $\boldsymbol{\Phi}$ as

$$\bar{\boldsymbol{N}} = f_{\text{ps2}}(\boldsymbol{\Phi}; \boldsymbol{\theta}_{\text{ps2}}). \tag{9}$$

Here, f_{ps1} is a feed-forward CNN of three layers with learnable parameters $\boldsymbol{\theta}_{\text{ps1}}$, where each layer applies 3x3 Conv of D_{ps} channels, BatchNorm (Ioffe & Szegedy, 2015), and ReLU. We use channels of $D_{\text{ps}} = 384$, and use no skip-connections or pooling. Similarly, f_{ps2} applies 3x3 Conv and L2 normalization that makes each $\bar{\boldsymbol{N}}_p$ a unit vector.

3.1.2. IMAGE RECONSTRUCTION NETWORK

IRNet synthesizes each observed image $\boldsymbol{I}_i$ as $\hat{\boldsymbol{I}}_i$ based on the rendering equation of Eq. (1). Specifically, IRNet first predicts $R = s\rho(\bar{\boldsymbol{n}}, \bar{\boldsymbol{\ell}}, \bar{\boldsymbol{v}})$, the multiplication of a cast shadow and a BRDF, under a particular illumination $\boldsymbol{\ell}_i$ as

$$\hat{\boldsymbol{R}}_i = f_{\text{ir}}(\boldsymbol{I}_i, \bar{\boldsymbol{N}}, \bar{\boldsymbol{\ell}}_i, \bar{\boldsymbol{v}}, \boldsymbol{\Phi}; \boldsymbol{\theta}_{\text{ir}}). \tag{10}$$

Here, we call $\hat{\boldsymbol{R}}_i \in \mathbb{R}^{C \times H \times W}$ a reflectance image, which is produced by a CNN f_{ir} as explained later. Then, IRNet synthesizes each image $\hat{\boldsymbol{I}}_i$ by the rendering equation below.

$$\hat{\boldsymbol{I}}_i = \hat{\boldsymbol{R}}_i \odot \max(\boldsymbol{\ell}_i^T \bar{\boldsymbol{N}}, \boldsymbol{0}) \tag{11}$$

Here, the inner products between light $\boldsymbol{\ell}_i$ and normal vectors $\bar{\boldsymbol{N}}$ are computed at each pixel p by $\boldsymbol{\ell}_i^T \bar{\boldsymbol{N}}_p$. Note that when $\boldsymbol{\ell}_i$ has color channels, we multiply a matrix $\boldsymbol{\ell}_i^T \in \mathbb{R}^{C \times 3}$ to $\bar{\boldsymbol{N}}_p$. Consequently, $\boldsymbol{\ell}_i^T \bar{\boldsymbol{N}}$ and $\hat{\boldsymbol{R}}_i$ have the same dimensions with $\boldsymbol{I}$. The $\max(\cdot, \boldsymbol{0})$ is done elementwise and is implemented by ReLU, and $\odot$ is elementwise multiplication. We now explain details of f_{ir} by dividing it into three parts.

Individual observation transform. The first part transforms each observed image $\boldsymbol{I}_i$ (which we denote as $\boldsymbol{X}_i$) into a feature map $\boldsymbol{Y}_i \in \mathbb{R}^{D_{\text{ir}} \times H \times W}$ as follows.

$$\boldsymbol{Y}_i = f_{\text{ir1}}(\boldsymbol{X}_i; \boldsymbol{\theta}_{\text{ir1}}) \tag{12}$$

The network architecture of f_{ir1} is the same with f_{ps1} in Eq. (8), except that we use channels of $D_{\text{ir}} = 16$ for f_{ir1}. To more effectively learn BRDFs, we use an additional specularity channel $\boldsymbol{S}_i$ for the input $\boldsymbol{X}_i$ as

$$\boldsymbol{X}_i = \text{Concat}(\boldsymbol{I}_i, \boldsymbol{S}_i), \tag{13}$$

where $\boldsymbol{S}_i \in \mathbb{R}^{1\times H\times W}$ is computed at each pixel p as

$$\boldsymbol{S}_{ip} = \bar{\boldsymbol{v}}^T \bar{\boldsymbol{s}}_{ip} = \bar{\boldsymbol{v}}^T \left[2(\bar{\boldsymbol{\ell}}_i^T \bar{\boldsymbol{N}}_p)\bar{\boldsymbol{N}}_p - \bar{\boldsymbol{\ell}}_i\right]. \tag{14}$$

Here, $\bar{\boldsymbol{s}}_{ip}$ is the direction of the specular reflection (dashed line between $\bar{\boldsymbol{n}}$ and $\bar{\boldsymbol{v}}$ in Fig. 2). It is well known by past studies that $\boldsymbol{S}_{ip}$ is highly correlated with the actual specular component of a BRDF. Therefore, directly giving it as a hint to the network will promote learning of complex BRDFs.

Global observation blending. Because $\boldsymbol{Y}_i$ has limited observation information under a particular illumination $\boldsymbol{\ell}_i$, we enrich it by $\boldsymbol{\Phi}$ in Eq. (8) that has more comprehensive information of the scene. We do this similarly to global and local feature blending in (Charles et al., 2017; Iizuka et al., 2016) as

$$\boldsymbol{Z}_i = f_{\text{ir2}}(\text{Concat}(\boldsymbol{Y}_i, \boldsymbol{\Phi}); \boldsymbol{\theta}_{\text{ir2}}), \tag{15}$$

where f_{ir2} applies 1x1 Conv, BatchNorm, and ReLU. Note that applying Conv to $\text{Concat}(\boldsymbol{Y}_i, \boldsymbol{\Phi})$ is efficiently done as $\boldsymbol{W}_1\boldsymbol{Y}_i + \boldsymbol{W}_2\boldsymbol{\Phi} + \boldsymbol{b}$ where Conv of $\boldsymbol{W}_2\boldsymbol{\Phi} + \boldsymbol{b}$ is computed only once and reused for all observations i.

Output. After the blending, we finally output $\hat{\boldsymbol{R}}_i$ by

$$\hat{\boldsymbol{R}}_i = f_{\text{ir3}}(\boldsymbol{Z}_i; \boldsymbol{\theta}_{\text{ir3}}), \tag{16}$$

where f_{ir3} is 3x3 Conv, BatchNorm, ReLU, and 3x3 Conv. As explained in Eq. (11), the resulting $\hat{\boldsymbol{R}}_i$ is used to reconstruct each image as $\hat{\boldsymbol{I}}_i$, which is the final output of IRNet.

Note that the internal channels of IRNet are all the same as D_{ir}. Also, IRNet simultaneously reconstructs all images during SGD iterations, by treating them as a minibatch: $\hat{\boldsymbol{I}} = \text{Batch}(\hat{\boldsymbol{I}}_1, \hat{\boldsymbol{I}}_2, \ldots, \hat{\boldsymbol{I}}_M) \in \mathbb{R}^{M\times C\times H\times W}$. This learning procedure is more explained in the next section.

3.2. Learning procedures (optimization)

We optimize the network parameters $\boldsymbol{\theta}$ by minimizing the following loss function using SGD.

$$L = L_{\text{rec}}(\hat{\boldsymbol{I}}, \boldsymbol{I}) + \lambda_t L_{\text{prior}}(\bar{\boldsymbol{N}}, \bar{\boldsymbol{N}}') \tag{17}$$

The first term defines reconstruction loss between the synthesized $\hat{\boldsymbol{I}}$ and observed images $\boldsymbol{I}$, which is explained in Sec. 3.2.1. The second term defines weak supervision loss between the predicted $\bar{\boldsymbol{N}}$ and some prior normal map $\bar{\boldsymbol{N}}'$. This term is only activated in early iterations of SGD (*i.e.*, $\lambda_t = 0$ when $t > T$) in order to warm up randomly initialized networks and stabilize the learning. This is more explained in Sec. 3.2.2. Other implementation details and hyper-parameter settings are described in Sec. 3.2.3.

Most importantly, the network is directly fit for a particular test scene without any pre-training on other data, by updating the network parameters $\boldsymbol{\theta}$ over SGD iterations. Final results are obtained at convergence.

3.2.1. Reconstruction loss

The reconstruction loss is defined as mean absolute errors between $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}$ over target object regions $\boldsymbol{O}$ as

$$L_{\text{rec}}(\hat{\boldsymbol{I}}, \boldsymbol{I}) = \frac{1}{MCO}\sum_{i,c,p} \boldsymbol{O}_p |\hat{\boldsymbol{I}}_{icp} - \boldsymbol{I}_{icp}|. \tag{18}$$

Here, $\boldsymbol{O}$ in $\{1,0\}^{1\times H\times W}$ is the binary object mask, and $O = \sum_p \boldsymbol{O}_p$ is its object area size. Using absolute errors increases the robustness to high-intensity specular highlights.

3.2.2. Early-stage weak supervision

If the target scene has relatively simple reflectance properties, the reconstruction loss alone can often lead to a good solution, even starting with randomly initialized networks. However, for complex scenes, we need to warm up the network by adding the following weak supervision.

$$\lambda_t L_{\text{prior}}(\bar{\boldsymbol{N}}, \bar{\boldsymbol{N}}') = \lambda_t \frac{1}{O}\sum_p \boldsymbol{O}_p \|\bar{\boldsymbol{N}}_p - \bar{\boldsymbol{N}}'_p\|_2^2 \tag{19}$$

Here, the prior normal map $\bar{\boldsymbol{N}}'$ is obtained by the simplest least squares method described in Sec. 2.1 using all observed pixels without any thresholding. Due to the presence of shadows and non-Lambertian specularity, this least squares solution can be very inaccurate. However, even such priors work well in our method, because we only use them to guide the optimization in its early stage. For this, we set λ_t to $0.1c$ for initial 50 iterations, and then set it to zero afterwards. The coefficient c is to adaptively balance weights between L_{rec} and L_{prior}, and is computed as the mean intensities of $\boldsymbol{I}$ over target object regions, *i.e.*, $c = L_{\text{rec}}(\boldsymbol{0}, \boldsymbol{I})$.

3.2.3. Implementation details

We use Adam (Kingma & Ba, 2015) as the optimizer. For each test scene, we iterate SGD updates for 1000 steps. Adam's hyper-parameter α is set to $\alpha_0 = 8 \times 10^{-4}$ for first 900 iterations, and then decreased to $\alpha_0/10$ for last 100 iterations for fine-tuning. We use the default values for the other hyper-parameters. The convolution weights are randomly initialized by He initialization (He et al., 2015).

In each iteration, PSNet predicts a surface normal map $\bar{\boldsymbol{N}}$, and then IRNet reconstructs all observed images $\hat{\boldsymbol{I}}$ as samples of a minibatch. Given $\bar{\boldsymbol{N}}$ and $\hat{\boldsymbol{I}}$, we compute the loss L and update the parameters $\boldsymbol{\theta}$ of both networks.

When computing the reconstruction loss L_{rec} in Eq. (18), we randomly dropout 90% of its elements and rescale L_{rec} by a factor of 10 instead. This treatment is to compensate for the well known issue of poor local convergence of SGD by the use of a large minibatch (Keskar et al., 2017).

Because we learn network parameters during testing, we always run BatchNorm by the training mode using statistics of given data (*i.e.*, we never use moving-average statistics).

Before being fed into the network, the input images $\boldsymbol{I}$ are cropped by a loose bounding box of the target object regions for reducing redundant computations. Then, the images are normalized by global scaling as

$$\boldsymbol{I}' = \boldsymbol{I}/(2\sigma), \tag{20}$$

where σ is the square-root of mean squared intensities of $\boldsymbol{I}$ over target regions. For PSNet, the normalized image tensor $\boldsymbol{I}'$ is further concatenated with the binary mask $\boldsymbol{O}$ as input.

4. Experiments

In this section we evaluate our method using a challenging real-world scene benchmark called DiLiGenT (Shi et al., 2018). In Sec. 4.1, we show comparisons with state-of-the-art photometric stereo methods. We then more analyze our network architecture in Sec. 4.2 and weak supervision technique in Sec. 4.3. In the experiments, we use $M = 96$ of observed images for each scene provided by the DiLiGenT dataset. Our method is implemented in Chainer (Tokui et al., 2015) and is run on a single nVidia Tesla V100 GPU with 16 GB memory and 32 bit floating-point precision.

4.1. Real-world scene benchmark (DiLiGenT)

We show our results on the DiLiGenT benchmark (Shi et al., 2018) in Table 1, where we compare our method with ten existing methods by mean angular errors. We also show visual comparisons of the top three and baseline methods for READING and HARVEST in Fig. 4. Our method achieves the best average score and best individual scores for eight scenes (excepting only two scenes of GOBLET and HARVEST) that contain various materials and reflectance surfaces. This is remarkable considering that another neural network method (Santo et al., 2017) outperforms the other existing methods only for HARVEST, in spite of its supervised learning. This HARVEST is the most difficult scene of all due to heavy interactions of cast shows and interreflections as well as spatially-varying materials and complex metallic BRDFs. For such complex scenes, supervised pre-training (Santo et al., 2017) is effective. The baseline method poorly performs especially for specular objects. Although we use its results as guidance priors, its low accuracy is not critical to our method thanks to the proposed early-stage supervision. We more analyze it in Sec. 4.3.

4.2. Analysis of the network architecture

In the middle part of Table 2, we show performance changes of our method by modifying its architecture. Specifically, we test two settings where we disable two connections from PSNet to IRNet, *i.e.*, the specularity channel input and the global observation blending described in Sec. 3.1.2. As shown, the proposed full architecture performs best, while the removal of the specularity channel input has the most negative impact. As expected, directly inputting a specularity channel indeed eases learning of complex BRDFs (*e.g.*, metallic surfaces in COW), demonstrating a strength of our physics-based network architecture that can exploit known pysical reflectance properties for BRDF learning.

4.3. Effects of early-stage weak supervision

We here evaluate the effectiveness of our learning strategy using early-stage weak supervision, by comparing with two cases where we use no or all-stage supervision (*i.e.*, λ_t is 0 or constant). See the bottom part of Table 2 for performance comparisons. Learning with no supervision produces comparable median scores but worse mean scores, compared to early-stage supervision. This indicates that learning with no supervision is very unstable and often gets stuck at bad local minimums, as shown in Fig. 5 (green profiles). On the other hand, learning with all-stage supervision is relatively stable but is strongly biased by inaccurate least-squares priors, often producing worse solutions as shown in Fig. 5 (blue profiles). In contrast, learning with the proposed early-stage supervision (red profiles) is more stable and persistently continues to improve accuracy even after terminating the supervision at $t = 50$ (shown as vertical dashed lines).

5. Discussions and related work

Our method is inspired by recent work of *deep image prior* by Ulyanov et al. (2018). They show that architectures of CNNs themselves behave as good regularizers for natural images, and show successful results for unsupervised tasks such as image super-resolution and inpainting by fitting a CNN for a single test image. However, their simple glass-hour network does not directly apply to photometric stereo, because we here need to simultaneously consider surface normal estimation that accounts for global statistics of observations, as well as reconstruction of individual observations for defining the loss. Our novel architecture addresses this problem by resorting to ideas of classical physics-based approaches to photometric stereo.

Our network architecture is also partly influenced by that of (Santo et al., 2017), which regresses per-pixel observations $\boldsymbol{I}_p \in \mathbb{R}^M$ to a 3D normal vector using a simple feed-forward network of five fully-connected and ReLU layers plus an output layer. Our PSNet becomes similar to theirs, if we use

Table 1. **Comparisons on ten real-world scenes of the DiLiGenT photometric stereo benchmark (Shi et al., 2018).** We compare our proposed method with ten existing calibrated photometric stereo methods. Here, we show mean angular errors in degrees (*i.e.*, the mean of $\arccos(\bar{\boldsymbol{N}}_p^T \bar{\boldsymbol{N}}_p^\star)$ over the object regions using ground truth normals $\bar{\boldsymbol{N}}_p^\star$) for ten scenes, and average scores. Our method achieves best accuracies for all except two scenes of GOBLET and HARVEST. The second best method (Santo et al., 2017) also uses a deep neural network, but it requires supervised pre-training on synthetic data and outperforms the other existing methods only for HARVEST. The results of the baseline least squares method are used in our method as prior normals for weak supervision. Since the priors are used only for an early-stage of learning, their low accuracies are not critical to the performance of our method. Note that, due to a non-deterministic property of our method, its accuracy for each scene is evaluated as the median score of 11 rounds run.

	BALL	CAT	POT1	BEAR	POT2	BUDDHA	GOBLET	READING	COW	HARVEST	AVG.
Proposed	**1.47**	**5.44**	**6.09**	**5.79**	**7.76**	**10.36**	11.47	**11.03**	**6.32**	22.59	**8.83**
Santo et al. (2017)	2.02	6.54	7.05	6.31	7.86	12.68	11.28	15.51	8.01	**16.86**	9.41
Shi et al. (2014)	1.74	6.12	6.51	6.12	8.78	10.60	10.09	13.63	13.93	25.44	10.30
Ikehata & Aizawa (2014)	3.34	6.74	6.64	7.11	8.77	10.47	**9.71**	14.19	13.05	25.95	10.60
Goldman et al. (2010)	3.21	8.22	8.53	6.62	7.90	14.85	14.22	19.07	9.55	27.84	12.00
Alldrin et al. (2008)	2.71	6.53	7.23	5.96	11.03	12.54	13.93	14.17	21.48	30.50	12.61
Higo et al. (2010)	3.55	8.40	10.85	11.48	16.37	13.05	14.89	16.82	14.95	21.79	13.22
Wu et al. (2010)	2.06	6.73	7.18	6.50	13.12	10.91	15.70	15.39	25.89	30.01	13.35
Ikehata et al. (2012)	2.54	7.21	7.74	7.32	14.09	11.11	16.25	16.17	25.70	29.26	13.74
Shi et al. (2012)	13.58	12.34	10.37	19.44	9.84	18.37	17.80	17.17	7.62	19.30	14.58
Baseline (least squares)	4.10	8.41	8.89	8.39	14.65	14.92	18.50	19.80	25.60	30.62	15.39

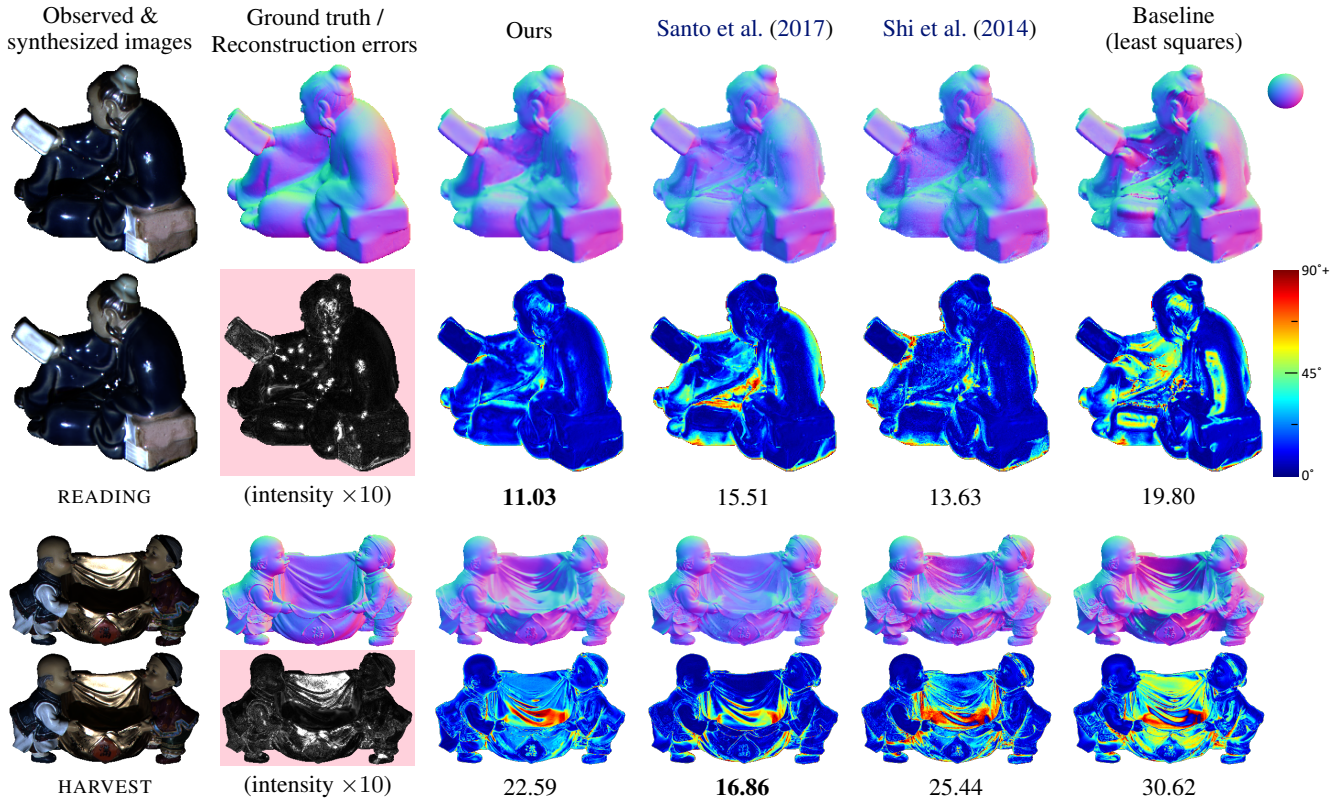
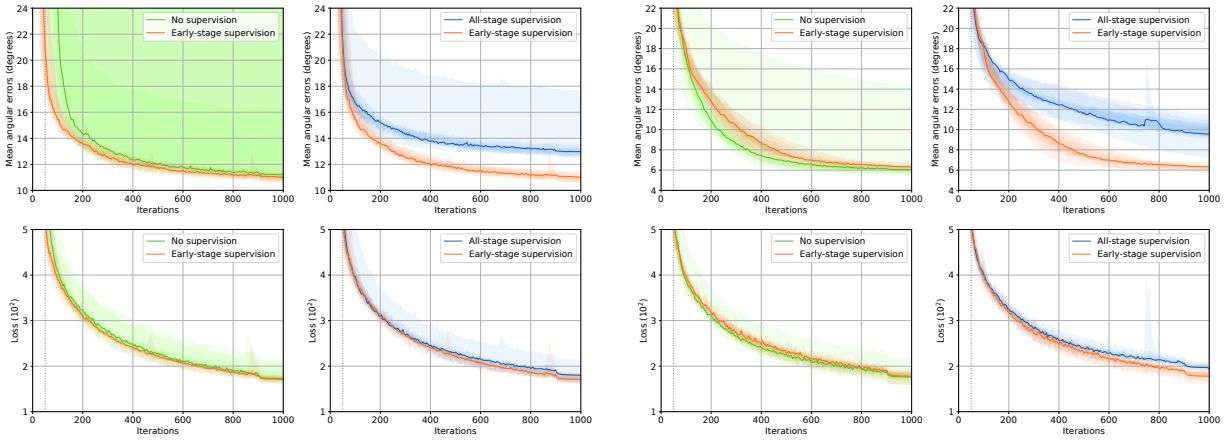


Figure 4. **Visual comparisons for READING and HARVEST scenes.** From left to right columns in each scene, we show 1) observed and our synthesized images, 2) ground truth normal and image reconstruction error maps, and 3–6) estimated surface normal and angular error maps by four methods. Numbers under angular error maps show their mean errors. See the supplementary material for more comparisons.

Table 2. **Evaluations of the proposed network architecture and weak supervision.** For each item we show median and mean scores (in left and right) by 11 rounds run. Here, **S**, **G** and **WS** denote the specularity input, global observation blending, and weak supervision by a prior normal map, respectively. Cell colors of red/blue indicate worse/better relative accuracy compared to the proposed settings.

	S	G	WS	BALL		CAT		POT1		BEAR		POT2		BUDDHA		GOBLET		READING		COW		HARVEST		AVG.	
Proposed	✓	✓	✓	**1.47**	**1.50**	5.44	**5.38**	**6.09**	**6.15**	5.79	5.84	7.76	7.71	10.36	10.22	11.47	**11.35**	11.03	10.98	6.32	6.26	22.59	**22.63**	**8.83**	**8.80**
No specular input		✓	✓	1.64	1.63	7.09	7.06	7.78	7.77	5.53	**5.55**	8.47	8.34	11.23	11.22	14.53	14.59	**10.71**	**10.75**	19.04	18.83	26.75	26.71	11.28	11.25
No global observation	✓		✓	1.50	1.50	13.18	15.12	8.47	8.50	5.76	5.74	**7.50**	**7.51**	12.76	12.68	12.50	12.54	16.81	20.20	**5.40**	**5.44**	25.12	25.34	10.90	11.46
No supervision	✓	✓		1.61	1.58	**5.30**	5.97	6.25	10.91	**5.53**	8.10	8.18	8.70	**10.08**	**10.16**	11.67	14.29	11.20	20.27	6.03	6.72	**22.48**	32.12	**8.83**	11.88
All-stage supervision	✓	✓	★	1.65	1.63	5.50	5.55	6.20	6.16	5.56	5.55	8.12	8.12	10.18	10.22	**11.34**	11.54	12.98	13.37	9.56	9.80	24.05	23.90	9.51	9.58



READING (early-stage vs. no/all-stage supervision) COW (early-stage vs. no/all-stage supervision)

Figure 5. **Convergence analysis with different types of weak supervision.** We show learning curves of mean angular errors (top) and loss values (bottom) for READING and COW, profiled by distributions of 11 rounds run (colored region) and medians (solid line). Compared to the proposed early-stage supervision (red), using no/all-stage supervision (green/blue) is often unstable or inaccurate. Vertical lines at $t = 50$ indicate termination of early-stage supervision. See the supplementary material for results of other scenes. Best viewed in color.

1x1 Conv with more layers and channels (*i.e.*, they use channels of 4096 and 2048 for the five internal layers). Since our method only needs to learn reflectance properties of a single test scene, our PSNet requires fewer layers and channels. More importantly, we additionally introduce IRNet, which allows direct unsupervised learning on test data.

There are some other early studies on photometric stereo using (shallow) neural networks. These methods work under more restricted conditions, *e.g.*, assuming pre-training by a calibration sphere of the same material with target objects (Iwahori et al., 1993; 1995), special image capturing setups (Iwahori et al., 2002; Ding et al., 2009), or the Lambertian surfaces (Cheng, 2006; Elizondo et al., 2008), whereas none of them is required by our method.

Currently, our method has limitations of a slow running time (*e.g.*, 1 hour to do 1000 SGD iterations for each scene) and limited performances to complex scenes (*e.g.*, HARVEST). However, several studies (Akiba et al., 2017; You et al., 2017; Goyal et al., 2017) show fast training of CNNs using extremely large minibatches and tuned scheduling of SGD step-sizes. Since our dense prediction method can use at most a large minibatch of $M \times H \times W$ pixel samples, the use of such acceleration schemes may improve the convergence speed. Also, a pre-training approach similar to (Santo et al., 2017) is still feasible for our method, which will accelerate the convergence and will also increase accuracy to complex scenes (with the loss of permutation invariance). Thorough analyses in such directions are left as our future work.

6. Conclusions

In this paper, we have presented a novel CNN architecture for photometric stereo. The proposed unsupervised learning approach bridges a gap between existing supervised neural network methods and many other classical physics-based unsupervised methods. Consequently, our method can learn complicated BRDFs by leveraging both powerful expressibility of deep neural networks and physical reflectance properties known by past studies, achieving the state-of-the-art performance in an unsupervised fashion just like classical methods. We also hope that our idea of physics-based unsupervised learning stimulates further research on tasks that lack of ground truth data for training, because even so the physics is everywhere in the real world, which will provide strong clues for the hidden data we desire.

Acknowledgements

The authors would thank Shi et al. (2018) for building a photometric stereo benchmark, Santo et al. (2017) for providing us their results, and Profs. Yoichi Sato and Ryo Yonetani and anonymous reviewers for their helpful feedback. The authors would gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU.

References

Akiba, T., Suzuki, S., and Fukuda, K. Extremely Large Minibatch SGD: Training ResNet-50 on ImageNet in 15 Minutes. *CoRR*, abs/1711.04325, 2017.

Alldrin, N., Zickler, T., and Kriegman, D. Photometric stereo with non-parametric and spatially-varying reflectance. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2008.

Belhumeur, P. N., Kriegman, D. J., and Yuille, A. L. The Bas-Relief Ambiguity. *Int'l J. Comput. Vis. (IJCV)*, 35 (1):33–44, Nov 1999.

Charles, R. Q., Su, H., Kaichun, M., and Guibas, L. J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 77–85, 2017.

Cheng, W.-C. Neural-network-based photometric stereo for 3d surface reconstruction. In *Proc. IEEE Int'l Joint Conf. Neural Network*, pp. 404–410, 2006.

Chung, H.-S. and Jia, J. Efficient photometric stereo on glossy surfaces with wide specular lobes. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2008.

Ding, Y., Iwahori, Y., Nakamura, T., Woodham, R. J., He, L., and Itoh, H. Self-calibration and Image Rendering Using RBF Neural Network. In *Proc. Int'l Conf. Knowledge-Based and Intell. Inf. and Engin. Syst.*, pp. 705–712, 2009.

Elizondo, D., Zhou, S.-M., and Chrysostomou, C. Surface Reconstruction Techniques Using Neural Networks to Recover Noisy 3D Scenes. In *Proc. Int'l Conf. Artificial Neural Networks*, pp. 857–866, 2008.

Esteban, C. H., Vogiatzis, G., and Cipolla, R. Multiview photometric stereo. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 30(3):548–554, 2008.

Furukawa, Y. and Ponce, J. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 32(8):1362–1376, 2010.

Georghiades, A. S. Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo. In *Proc. Int'l Conf. Comput. Vis. (ICCV)*, pp. 816–823, 2003.

Goldman, D. B., Curless, B., Hertzmann, A., and Seitz, S. M. Shape and Spatially-Varying BRDFs from Photometric Stereo. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 32(6):1060–1071, 2010.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *CoRR*, abs/1706.02677, 2017.

He, H., Xin, B., Ikehata, S., and Wipf, D. P. From Bayesian Sparsity to Gated Recurrent Nets. In *Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 5560–5570. 2017a.

He, K., Zhang, X., Ren, S., and Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proc. Int'l Conf. Comput. Vis. (ICCV)*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *Proc. Int'l Conf. Comput. Vis. (ICCV)*, 2017b.

Higo, T., Matsushita, Y., and Ikeuchi, K. Consensus photometric stereo. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 1157–1164, 2010.

Hold-Geoffroy, Y., Gotardo, P. F. U., and Lalonde, J. Deep Photometric Stereo on a Sunny Day. *CoRR*, abs/1803.10850, 2018.

Iizuka, S., Simo-Serra, E., and Ishikawa, H. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Trans. Graphics (ToG)*, 35(4):110:1–110:11, 2016.

Ikehata, S. and Aizawa, K. Photometric Stereo Using Constrained Bivariate Regression for General Isotropic Surfaces. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 2187–2194, 2014.

Ikehata, S., Wipf, D., Matsushita, Y., and Aizawa, K. Robust photometric stereo using sparse regression. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 318–325, 2012.

Ioffe, S. and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proc. Int'l Conf. Mach. Learn. (ICML)*, volume 37, pp. 448–456, 2015.

Iwahori, Y., Woodham, R. J., Tanaka, H., and Ishii, N. Neural network to reconstruct specular surface shape from its three shading images. In *Proc. 1993 Int'l Conf. Neural Networks*, volume 2, pp. 1181–1184, 1993.

Iwahori, Y., Bagheri, A., and Woodham, R. J. Neural network implementation of photometric stereo. In *Proc. Vision Interface*, 1995.

Iwahori, Y., Watanabe, Y., Woodham, R. J., and Iwata, A. Self-calibration and neural network implementation of photometric stereo. In *Proc. Int'l Conf. Pattern Recognit. (ICPR)*, volume 4, pp. 359–362, 2002.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., and Bry, A. End-To-End Learning of Geometry and Context for Deep Stereo Regression. In *Proc. Int'l Conf. Comput. Vis. (ICCV)*, 2017.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *Proc. Int'l Conf. Learn. Repres. (ICLR)*, 2017.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *Proc. Int'l Conf. Learn. Repres. (ICLR)*, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 1097–1105. 2012.

Matusik, W., Pfister, H., Brand, M., and McMillan, L. Efficient Isotropic BRDF Measurement. In *Proc. Eurographics Workshop on Rendering*, 2003.

Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 4040–4048, 2016.

Nehab, D., Rusinkiewicz, S., Davis, J., and Ramamoorthi, R. Efficiently Combining Positions and Normals for Precise 3D Geometry. *ACM Trans. Graphics (ToG)*, 24(3):536–543, 2005.

Park, J., Sinha, S. N., Matsushita, Y., Tai, Y. W., and Kweon, I. S. Robust Multiview Photometric Stereo using Planar Mesh Parameterization. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 39(8):1591–1604, 2017.

Santo, H., Samejima, M., Sugano, Y., Shi, B., and Matsushita, Y. Deep Photometric Stereo Network. In *Proc. Int'l Workshop Physics Based Vision meets Deep Learning (PBDL) in ICCV*, 2017.

Shi, B., Tan, P., Matsushita, Y., and Ikeuchi, K. Elevation Angle from Reflectance Monotonicity: Photometric Stereo for General Isotropic Reflectances. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 455–468, 2012.

Shi, B., Tan, P., Matsushita, Y., and Ikeuchi, K. Bi-Polynomial Modeling of Low-Frequency Reflectances. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 36(6): 1078–1091, 2014.

Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S.-K., and Tan, P. A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2018. (to appear).

Taniai, T., Matsushita, Y., Sato, Y., and Naemura, T. Continuous 3D Label Stereo Matching using Local Expansion Moves. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2017. (accepted).

Tokui, S., Oono, K., Hido, S., and Clayton, J. Chainer: a next-generation open source framework for deep learning. In *Proc. Workshop Mach. Learn. Syst. (LearningSys) in NIPS*, 2015. URL https://chainer.org.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. Deep Image Prior. In *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2018. (to appear).

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

Wu, L., Ganesh, A., Shi, B., Matsushita, Y., Wang, Y., and Ma, Y. Robust photometric stereo via low-rank matrix completion and recovery. In *Proc. Asian Conf. Comput. Vis. (ACCV)*, pp. 703–717, 2010.

Xin, B., Wang, Y., Gao, W., Wipf, D. P., and Wang, B. Maximal Sparsity with Deep Networks? In *Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 4340–4348. 2016.

You, Y., Zhang, Z., Hsieh, C., and Demmel, J. ImageNet Training in Minutes. *CoRR*, abs/1709.05011, 2017.