# Learning Inter- and Intraframe Representations for Non-Lambertian Photometric Stereo

Yanlong Cao<sup>a,b</sup>, Binjie Ding<sup>a,b</sup>, Zewei He<sup>a,b</sup>, Jiangxin Yang<sup>a,b</sup>, Jingxi Chen<sup>a,b</sup>, Yanpeng Cao<sup>a,b,\*</sup>, Xin Li<sup>c</sup>

<sup>a</sup>State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China <sup>b</sup>Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China <sup>c</sup>School of the Electrical Engineering and Computer Science (EECS), Louisiana State University, Baton Rouge, LA 70803, USA

#### **Abstract**

Photometric stereo provides an important method for high-fidelity 3D reconstruction based on multiple intensity images captured under different illumination directions. In this paper, we present a complete framework, including a multilight source illumination and acquisition hardware system and a two-stage convolutional neural network (CNN) architecture, to construct inter- and intraframe representations for accurate normal estimation of non-Lambertian objects. We experimentally investigate numerous network design alternatives for identifying the optimal scheme to deploy inter- and intraframe feature extraction modules for the photometric stereo problem. Moreover, we propose utilizing the easily obtained object mask to eliminate adverse interference from invalid background regions in intraframe spatial convolutions, thus effectively improving the accuracy of normal estimation for surfaces made of dark materials or with cast shadows. Experimental results demonstrate that the proposed masked two-stage photometric stereo CNN model (MT-PS-CNN) performs favourably against state-of-the-art photometric stereo techniques in terms of both accuracy and efficiency. In addition, the proposed method is capable of predicting accurate and rich surface normal details for non-Lambertian objects of complex geometry and performs stably given inputs captured in both sparse and dense lighting distributions.

Keywords: Photometric Stereo, Convolutional Neural Network (CNN), 3D Reconstruction/Modeling

# 1. Introduction

In recent years, photometric stereo has received significant attention in the field of optical engineering and advanced manufacturing [1, 2]. Photometric stereo techniques estimate accurate and highly detailed surface normals of a target object based on a set of images captured in different light directions using a viewpoint fixed camera. These techniques can generate 3D models with rich details to facilitate various applications, such as automated industrial quality inspection [3], high-fidelity 3D reconstruction/modelling [4, 5], and face recognition and verification [6, 7].

The basic theory of photometric stereo was first proposed by Woodham et al. based on the assumption of ideal Lambertian reflectance [8]. However, most of the real-world objects are non-Lambertian. Therefore, many researchers have attempted to utilize flexible surface reflection functions and bidirectional reflection distribution functions (BRDFs) to develop more applicable photometric stereo techniques that work well for real-world objects [9, 10].

In recent years, numerous deep learning-based methods have been proposed for high-quality photometric stereo tasks. Some approaches have attempted to explore the per-pixel intensity

\*Corresponding author

Email address: caoyp@zju.edu.cn (Yanpeng Cao)

variation among different images (interframe clues [11]) to generate high-accuracy surface normal estimation results. However, these interframe photometric stereo methods neglect local spatial intensity variation among neighbouring pixels, encode important cues for predicting surface normals, and thus perform unsatisfactorily when the number of input images decreases. To overcome the limits of interframe photometric stereo techniques, some researchers took advantage of per-frame intensity variation among neighbouring pixels (intraframe clues [12, 13]) to robustly estimate the surface normals when a limited number of input images were available. It is worth noting that the intraframe photometric stereo technique typically performs unfavourably compared with those methods built on interframe analysis. Therefore, it is highly desirable to develop a unified method that performs both inter- and intraframe analyses for the challenging surface normal estimation task.

In this paper, we present a complete photometric stereo data acquisition and processing framework, as shown in Fig. 1, constructing inter- and intraframe feature representations based on an arbitrary number of unordered images captured under different lighting configurations for high-quality surface normal estimation of non-Lambertian objects. The image acquisition system is installed on the end of a robot manipulator for image capture of target objects of various sizes. Sixteen LED lights are evenly distributed around a viewpoint-fixed camera to illuminate the target object from different directions for im-

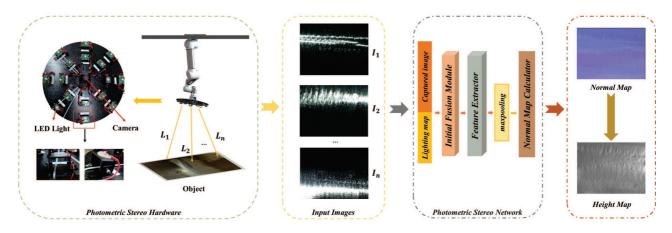


Fig. 1. The data acquisition and processing workflow of our proposed photometric stereo system.

age capture. Given an arbitrary number of images captured under different lighting configurations, we further present a novel two-stage CNN architecture to explore valuable information encoded in both local image patches and cross adjacent frames and construct inter- and intraframe feature representations for high-quality surface normal estimation of non-Lambertian objects, as illustrated in Fig. 2. More specifically, we utilize  $M \times 1 \times 1$  convolutional layers to analyse lighting variations of individual pixels on M frames and  $1 \times N \times N$  convolutional layers to capture intraframe intensity variation among  $N \times N$  local pixels. We experimentally evaluate the performances of various network design alternatives in an attempt to identify the optimal sequence and strategy to deploy interframe and intraframe feature extraction modules. Two important findings are noted: (1) it is better to first perform interframe feature extraction followed by intraframe feature extraction since the frame-to-frame observations provide important information for the photometric stereo task; (2) it is desirable to divide the entire feature extraction process into two individual stages, whereas mixing the inter- and intraframe feature extraction steps will adversely affect the performance of surface normal estimation. Moreover, we propose utilizing the easily obtained object mask to eliminate adverse interference from invalid background regions in intraframe spatial convolutions, thus effectively improving the accuracy of normal estimation for surfaces with insufficient reflectance observations (e.g., made of highly absorptive dark materials or with cast shadows). Compared with the state-of-theart CNN-based photometric stereo techniques [12, 11], our proposed MT-PS-CNN is capable of estimating more accurate surface normals using fewer parameters and can perform consistently well across various image capturing configurations. This work provides the following three main contributions.

- We design a two-stage CNN architecture to construct interand intraframe representations for photometric stereo and establish ablation studies to identify the optimal scheme to deploy interframe and intraframe feature extraction modules to achieve high-quality surface normal estimation.
- We propose utilizing the easily obtained object mask to eliminate adverse interference from invalid background re-

- gions during intraframe spatial convolutions, which provides an effective technique to facilitate accurate normal estimation for surfaces made of highly absorptive dark materials or with cast shadows.
- With fewer parameters, our proposed MT-PS-CNN outperforms state-of-the-art photometric stereo techniques [12, 11]. Moreover, the proposed method is capable of predicting accurate and rich surface normal details for non-Lambertian objects and performs well with sparse input frames.

# 2. Related Work

In this section, we provide a brief overview of conventional and deep-learning based photometric stereo methods for non-Lambertian objects. For a detailed introduction of recent studies of photometric stereo, readers can refer to [14].

### 2.1. Conventional methods

The original photometric stereo method [8] works based on the ideal Lambertian reflectance model and analyses per-pixel lighting observation variations. However, most real-world objects cannot satisfy the ideal Lambertian reflectance model; thus, non-Lambertian photometric stereo methods have been extensively researched given their increased practicability. In general, existing non-Lambertian methods can be classified into three major categories.

The first category includes robust-estimation-based approaches, which treat the non-Lambertian reflectances as outliers. The methods assume that the majority of observations conform to the Lambertian model and that the non-Lambertian reflectances are sparse and local [15, 16]. Recently, studies have explored relevant approaches, e.g., rank-minimization-based methods [17], using RANSAC schemes [18] and taking median-based approaches [19]. Due to the limitation of the assumption, robust estimation-based methods require numerous inputs and experience difficulties with the thickset non-Lambertian surfaces, i.e., board shadowed areas or extensive specularity.

The second category includes reflectance model-based approaches, which arrange parametric or nonparametric models to precisely represent the appearance of real-world materials. These sophisticated reflectance models can be divided into physically based (e.g., Torrance-Sparrow model [20, 21, 22], Cook-Torrance model [23] and the Ward model [24, 25]) and empirical models (e.g., Phong model [26], Blinn-Phong model [27]). However, these sophisticated reflectance model-based methods are only applicable for targets made of limited classes of materials [28, 29].

The third category includes example-based methods, recovering the surface normal with an additional reference object. Hertzmann and Seitz proposed an example-based method [30] that uses orientation consistency to reconstruct the surface normals of the target object. Orientation consistency means that two points with the same surface orientation should have the same or similar appearance in an image. The example-based method can obtain the surface without solving a complex optimization; however, it requires a known-surface-parameter reference object that is typically not available during practical image acquisition.

# 2.2. Deep-learning-based methods

Due to the adovementioned limitations of conventional photometric stereo methods, deep learning techniques with the ability to approximate highly nonlinear mappings have been recently utilized to solve complex photometric stereo problems. Santo et al. [31] first solved the photometric stereo task via deep learning and proposed a pixelwise method that utilizes a fully connected network (DPSN) to establish a mapping from given observations to the surface normal. The assumption that light directions must be predefined and remain the same during the training and test phases restricts the application of DPSN. Ikehata et al. [11] introduced a CNN-based method, CNN-PS, which relaxes the limitation of lighting information and image structure through a novel network input (2-D observation map). A synthetic photometric stereo dataset, called Cycles PS, was presented in this paper, considering the global illumination effects. CNN-PS performs significantly better on the benchmark dataset than most traditional photometric stereo approaches. Chen et al. [12] first proposed a fully convolutional network called PS-FCN that also takes unstructured images and lighting information as input. PS-FCN takes the full-size images into account and effectively utilizes the spatial information among local pixels, and this process is neglected in the previously proposed pixelwise method. Thus, the frame-based PS-FCN method performs better with a sparse input setting than the pixelwise methods. Then, Chen et al. [13] extended PS-FCN and proposed an uncalibrated PS method (SDPS-Net) for non-Lambertian surfaces. SDPS-Net contains two processing steps: a classification network (LCNet) is used to approximate the light information, and a subsequent prediction network (NENet) is utilized to estimate the surface normal.

Given sufficient input images, pixelwise methods (e.g., CNN-PS [11] and DPSN [31]) are generally more accurate for surface normal estimations than framewise methods (e.g., PS-FCN [12] and SDPS-Net [13]). However, a sparse input setting

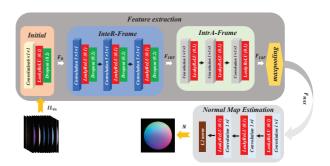


Fig. 2. The overall architecture of the proposed MT-PS-CNN model.

will adversely decrease the performance of pixelwise methods due given that these methods ignore spatial information. To handle sparse input, Li et al. [32] proposed a deep learning pixelwise approach, which applies a connection table that can select relatively effective light directions, to minify the input of pixelwise photometric stereo methods. Zheng et al. [33] proposed SPLINE-Net, which generates dense lighting observations by lighting interpolation, to improve the performance of sparse photometric stereo. Wang et al. [34] proposed a photometric stereo network that utilizes collocated light images as supplementary information to improve the performance.

## 3. Approach

Given q RGB images of a target object with  $p = W \times H$  pixels (W and H are the width and height of input images, respectively) captured under different light directions, surface normals of p pixels  $N \in \mathbb{R}^{3 \times p}$ , and light directions of q images  $L \in \mathbb{R}^{3 \times q}$ , the observation matrix  $I \in \mathbb{R}^{3 \times p \times q}$  can be formulated as follows[12]:

$$I = \Theta \circ \operatorname{repmat}(\max(N^{T}L, 0), 3),$$
 (1)

where  $\Theta \in \mathbb{R}^{3 \times p \times q}$  is a complex function of the surface normal, light direction, and viewing direction (the viewing direction is set to  $[0,0,1]^{\mathsf{T}}$ , which is parallel to the z-axis of the world coordinates), and  $\circ$  represents the elementwise dot product. In this paper, we present a two-stage CNN model MT-PS-CNN to extract inter- and intraframe feature representations to directly estimate the normal matrix N based on the observation matrix I and light direction matrix I without explicitly modelling the complex I0 function.

### 3.1. Network architecture

As illustrated in Fig. 2, our proposed MT-PS-CNN model consists of three major components: initial feature extraction, inter- and intraframe feature extraction, and normal map estimation. For each captured image, we follow conventional practice to replicate its corresponding 3-vector light direction along the spatial directions to obtain a  $3 \times H \times W$  lighting map [12]. We apply the binary object mask to the lighting map, extract light directions for the target object (mask pixel values equal to 1) and eliminate invalid background regions (mask pixel values equal to 0). More details on generating 2D masked lighting

maps are provided in Sec. 3.2. The obtained 3-channel masked lighting map is concatenated with the 3-channel RGB image to generate a 6-channel image-light data matrix. We stack q image-light matrices together to obtain the input of our MT-PS-CNN  $IL_{input}$  which is a  $(6 \times q) \times H \times W$  data matrix. Note that we purposely store input images and masked lighting masks in a 2D matrix for the following spatial convolutions.

Given input  $IL_{input}$ , we first deploy the initial feature extraction module to compute feature map  $F_0$  as follows:

$$F_0 = \mathcal{F}_{dropout}(\sigma(Con_{6\times 1\times 1}(IL_{input}))), \tag{2}$$

where  $Con_{6\times 1\times 1}(\cdot)$  denotes a 3D convolutional layer of the 6 × 1 × 1 kernel. Note that the kernel size of 3D convolution is set to 6 × 1 × 1 to process the concatenated 6-channel image-light data matrices. Following the 3D convolutional layer, a leaky ReLU layer  $\sigma(\cdot) = max(0.1x, x)$  and a dropout layer  $\mathcal{F}_{dropout}(\cdot)$  are deployed to activate the values and simulate the cast shadow effects [31, 11].

Then, we deploy a number of inter- and intraframe feature extraction blocks to perform simultaneous analysis of frame-to-frame and per-frame lighting variations. Within each InteR-frame Feature Extraction (IRFE) block, we utilize 3D convolutional layers of kernel size  $M \times 1 \times 1$  (i.e.,  $Con_{M \times 1 \times 1}(\cdot)$ ) to process lighting variations of individual pixels on M adjacent frames. Such interframe information provides important clues to eliminate the influence of outliers (i.e., shadows, interreflection, and specularity, etc.) and thus leads to accurate restoration results in photometric stereo tasks [35]. Each convolution is followed by a leaky ReLU activation and a dropout layer. We stack K IRFE blocks to compute interframe feature representations  $F_{IRF}^{K}$  as follows:

$$F_{IRF}^{1} = IRFE^{1}(F_0) \tag{3}$$

$$F_{IRF}^2 = IRFE^2 \left( F_{IRF}^1 \right) \tag{4}$$

:

$$F_{IRF}^{K} = IRFE^{K} \left( F_{IRF}^{N-1} \right), \tag{5}$$

where  $IRFE^{i}(\cdot) = \mathcal{F}_{dropout}(\sigma(Con^{i}_{M\times 1\times 1}(\cdot)))$  represents the operations of the  $i_{th}$  IRFE block.

The computed feature map  $F_{IRF}^{K}$  is then fed to a number of IntrA-frame feature extraction (IAFE) blocks to exploit the spatial information in local image patches and compute interframe feature representations  $F_{IAF}^{L}$  as follows:

$$F_{IAF}^{1} = IAFE^{1}\left(F_{IRF}^{K}\right) \tag{6}$$

$$F_{IAF}^2 = IAFE^2 \left( F_{IRF}^1 \right) \tag{7}$$

:

$$F_{IAF}^{L} = IAFE^{L}\left(F_{IRF}^{L-1}\right),\tag{8}$$

where L is the total number of stacked IAFE blocks, and  $IAFE^i(\cdot) = \sigma(Con^i_{1\times N\times N}(\cdot))$  denotes the operations of the  $i_{th}$  IAFE block. Note that each IAFE block contains 3D convolutional layers of kernel size  $1 \times N \times N$  (i.e.,  $Con_{1\times N\times N}(\cdot)$ ) and

leaky ReLU activation to capture intraframe intensity variation among  $N \times N$  local pixels. The extracted local context information can improve the performance of CNN models to handle various reflectances and performs robustly under sparse lighting distributions [36, 37]. In our implementation, we experimentally set K = L = 3 to achieve a good balance between model complexity and good performance.

To handle a flexible number of input images in photometric stereo tasks, order-agnostic operations (e.g., pooling layers) [38, 39] are typically utilized to standardize/fix the channel number of feature maps. Following the research work of Chen et al. [12], we apply the max-pooling operation (MP) to compress the channel number of  $F_{LAF}^{L}$  as follows:

$$F_{MAX} = MP(F_{IAF}^L), (9)$$

where  $F_{MAX}$  is the output of the max-pooling operation. It is a representation with a fixed number of channels by aggregating most salient features from images captured under different light directions.

A normal map estimation subnetwork is appended after the max-pooling operation for normal map estimation, converting the computed feature  $F_{MAX}$  to the surface normal  $N_{i,j}$  (i and j denote the spatial coordinates of the normal map) as follows:

$$N_{i,j} = NME(F_{MAX}),$$

$$= L_{norm}^2(\sigma(Con_{1\times 1}(\sigma(Con_{1\times 1}(F_{MAX})))))), \qquad (10)$$

where  $NEM(\cdot)$  denotes the operations of the Normal Map Estimation module, which contains three  $1 \times 1$  convolutional layers, two leaky ReLU activations and an L2-normalization layer  $L_{norm}^2$  for predicting the normal map.

The training of the proposed MT-PS-CNN model is driven by minimizing the error between the predicted and ground truth normal maps. We adopt the commonly used cosine similarity loss, which is formulated as follows:

$$\mathcal{L} = \frac{1}{HW} \sum_{i,j} (1 - N_{i,j} \cdot \tilde{N}_{i,j}), \tag{11}$$

where  $N_{i,j}$  and  $\tilde{N}_{i,j}$  denote the predicted and ground truth normal maps, respectively and  $\cdot$  indicates the dot product operation. When the predicted normal has a similar orientation as the ground truth,  $N_{i,j} \cdot \tilde{N}_{i,j}$  approaches 1, the loss approaches 0, and vice versa.

# 3.2. Masked lighting map

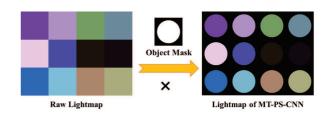


Fig. 3. Illustration of generating masked lighting maps.

In the photometric stereo task, large field-of-view cameras are typically utilized to cover the entire target object for image capturing. As a result, a large area of input image covers invalid backgrounds and contains pixels of unchanged RGB values. Note that these background pixels present similar lighting variation patterns as the object pixels with insufficient reflectance observations (e.g., surfaces made of highly absorptive dark materials or with cast shadows) and will cause confusion for surface normal inference, particularly in intraframe spatial convolutions. Therefore, we argue that it is important to exclude such invalid background pixels in the training/testing of CNN-based surface normal estimation models.

As a simple yet effective solution, we utilize the easily obtained object mask to extract pixels of the target object and eliminate invalid background regions. More specifically, we follow the conventional practice to replicate the 3-vector light direction of a captured image along the x and y directions to obtain a  $3 \times H \times W$  lighting map [12] and then apply the binary object mask (0 mask value defines a background pixel and 1 mask value defines an object pixel) to generate the masked lighting maps as illustrated in Fig. 3. In contrast to the existing uncalibrated photometric stereo method which concatenates object masks with individual captured images for light source estimation [40], we refer to the binary object mask to eliminate adverse interference from invalid background regions and thus achieve more accurate surface normal estimation results. We will experimentally evaluate the effectiveness of utilizing masked lighting maps to improve the accuracy of normal estimation for surfaces made of highly absorptive dark materials or with specular highlights and cast shadows.

# 4. Experimental Results

In this section, we systematically evaluate the performance of our proposed MT-PS-CNN and compare it with the state-of-theart photometric stereo methods on commonly used synthetic (MERL [41]) and real-world (DiLiGenT [42] and Light Stage Data Gallery [43]) datasets. We also validate the effectiveness of our proposed method using our own captured photometric stereo images of polished steel surfaces with tiny scratches.

# 4.1. Implementation details

All experiments were performed on a PC with GeForce GTX 1080Ti and 96 GB RAM. For training and testing, our model was implemented in PyTorch using 723 K learnable parameters. The Adam optimizer is used to optimize our networks with parameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use the *Blobby* and *Sculpture* datasets provided by [12] as our training data. There are 85212 samples on both datasets, where each sample contains 64 images rendered under randomly sampling light directions. The training process takes 30 epochs with a batch size of 32 (approximately 17 hours). We applied the same data augmentation technique as suggested in PS-FCN [12], except adding extra noise disturbances. The accuracy of surface normal estimation is quantitatively evaluated by computing the

mean angular error (MAE) between the predicted and ground truth normal maps as follows:

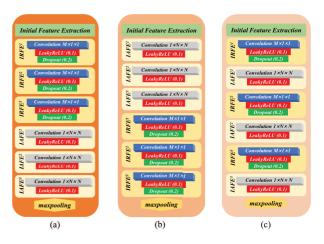
$$MAE = \arccos\left[\frac{1}{K}\sum_{k}(1 - N_k \cdot \tilde{N}_k)\right],\tag{12}$$

where  $N_k$  and  $\tilde{N}_k$  denote the predicted and ground truth normal maps, respectively, and K denotes the total number of target pixels in the normal map. Note that a lower MAE value indicates higher accuracy of normal estimation.

## 4.2. Performance Analysis

In this section, we establish ablation experiments to evaluate the effects of (1) the design of network architectures, (2) the incorporation of a masked lighting map, and (3) the setting of kernel sizes.

#### 4.2.1. Evaluation of network designs



**Fig. 4.** A number of network design alternatives for inter- and intraframe feature extraction. (a)T-IRFE-IAFE; (b)T-IAFE-IRFE; (c)M-IRFE-IAFE.

In this subsection, we discuss the best method to deploy interframe and intraframe feature extraction modules to achieve high-accuracy surface normal estimation and experimentally evaluate the performances of three network architectures as illustrated in Fig. 4.

The design of a CNN architecture to extract inter- and intraframe features for accurate surface normal estimation involves two critical design issues. The first issue is that both frame-to-frame (inter-) and per-frame (intra-) observations provide important information for the photometric stereo task and which feature extraction module should be deployed first. The second issue is whether the CNN architecture should be divided into two individual stages to perform inter- and intraframe feature extraction separately or mix the inter- and intraframe feature extraction steps.

Based on the two abovementioned critical design issues, we design three different network alternatives (Fig. 4) that employ different schemes to deploy inter- and intraframe feature extraction modules as follows:

(1) Two-stage IRFE-IAFE design (T-IRFE-IAFE): The first design incorporates a two-stage architecture by deploying

inter- and intraframe feature extraction steps in a cascaded manner. More specifically, it first deploys a number of IRFE blocks  $(IRFE^1 \Rightarrow IRFE^2 \dots \Rightarrow IRFE^K)$  to compute interframe features based on frame-to-frame lighting variations of individual pixels and then a number of IAFE blocks  $(IAFE^1 \Rightarrow IAFE^2 \dots \Rightarrow IAFE^L)$  to compute intraframe features by analysing intensity variation among spatial neighbouring pixels.

- (2) Two-stage IAFE-IRFE design (T-IAFE-IRFE): The second design also utilizes a two-stage architecture but switches the order to deploy inter- and intraframe feature extraction modules. Thus, it first deploys a number of IAFE blocks  $(IAFE^1 \Rightarrow IAFE^2 \dots \Rightarrow IAFE^L)$  and then a number of IRFE blocks  $(IRFE^1 \Rightarrow IRFE^2 \dots \Rightarrow IRFE^K)$  to compute features for surface normal estimation.
- (3) Mixed IRFE-IAFE design (M-IRFE-IAFE): Different from the above designs based on two-stage architecture, M-IRFE-IAFE deploys inter- and intraframe feature extraction steps in an alternative manner. More specifically, it makes use of a number of grouped IRFE and IAFE blocks ( $[IRFE^1, IAFE^1] \Rightarrow [IRFE^2, IAFE^2] \dots \Rightarrow [IRFE^K, IAFE^K]$ ) to compute inter- and intraframe features in different convolutional stages.

In our implementation, we experimentally set K = L = 3 to achieve a good balance between model complexity and good performance. For a fair comparison, these three different network design alternatives (T-IRFE-IAFE, T-IAFE-IRFE, and M-IRFE-IAFE) are trained and evaluated using the same parameters (M = 3, N = 3, K = 3, and L = 3). The comparative results (MAE) on the DiLiGenT dataset are shown in Table 1.

Table 1: Quantitative evaluation (MAE) of three different architectures (T-IRFE-IAFE, T-IAFE-IRFE, and M-IRFE-IAFE) to deploy inter- and intraframe feature extraction blocks on the DiLiGenT dataset.

Method	BALL	CAT	POT1	BEAR	POT2	BUDD.	GOBL.	READ.	COW	HARV.	Avg.
T-IRIA.	2.82	5.79	6.92	5.59	7.55	6.93	7.98	11.85	7.82	14.18	7.74
T-IAIR.	2.58	6.02	6.99	6.62	8.68	7.34	9.46	13.68	8.79	15.95	8.61
M-IRIA.	2.76	5.99	7.24	6.84	7.79	7.23	8.39	11.71	8.32	14.79	8.11

It is observed that the design of T-IRFE-IAFE significantly outperforms the design of T-IAFE-IRFE, and the average MAE of 10 objects is reduced by 0.87°. The comparative result shows that it is better to first perform interframe feature extraction and then intraframe feature extraction. This finding might be explained by the fact that although both inter- and intraframe features are useful for normal prediction, the interframe information is sensitive to outliers (specularity, shadow, etc.). Thus, it provides more important information to generate high-fidelity normal maps. In comparison, the intraframe features provide complementary information to efficiently eliminate the interference from outliers but adversely smooth out textured details. Our experimental results are consistent with previous research findings that pixelwise methods are generally more accurate for surface normal estimation than framewise methods if dense input images are provided [11, 31]. Therefore, it is reasonable to first extract interframe features that provide fundamental information for the photometric stereo task and then perform spatial

reasoning through intraframe feature extraction to further improve the accuracy of surface normal estimation.

Another interesting finding is that although M-IRFE-IAFE (altering inter- and intraframe feature extraction steps) is proven to be an effective design in many video sequence analysis tasks, such as gesture/action recognition [44], it is not very suitable for the photometric stereo task (M-IRFE-IAFE vs. T-IRFE-IAFE: 8.106° vs. 7.741°). Given that inter- and intraframe features provide two different types of information for estimating normal maps, it is more reasonable to divide the entire feature extraction process into two individual stages instead of mixing the inter- and intraframe feature extraction steps.

# 4.2.2. Effectiveness of masked lighting map

Table 2: Comparative results of various surface normal estimation models with/without referring to the masked lighting maps. For *Sphere* and *Bunny* objects, we calculate the average MAE values for 100 different materials.

Object	Method	M=3, N=1	M=1, N=3	M=3, N=3
Sphere	Mask	8.89	4.51	4.16
sphere	No-Mask	8.82	5.97	5.73
Dumm	Mask	7.93	4.39	4.03
Bunny	No-Mask	7.86	5.01	4.73

In this subsection, we experimentally evaluate the effectiveness of utilizing masked lighting maps to improve the accuracy of normal estimation. Here, we adopt the best-performing two-stage CNN architecture (T-IRFE-IAFE) for inter- and intraframe feature extraction and set kernel size parameters M (defining how many adjacent frames to process) and N (defining how many neighbouring pixels to process) to different values. Note that the model will only perform interframe convolutions by setting M=1 given that it only considers intensity variation among spatial neighbouring pixels on the current frame. Similarly, the model will become a per-pixel method and compute interframe features exclusively based on lighting variations of single pixels by setting N=1.

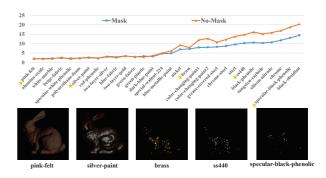
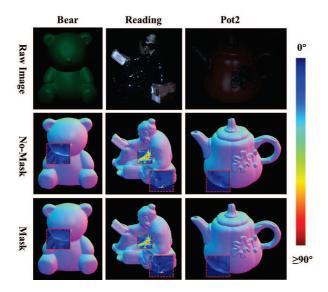


Fig. 5. Quantitative comparison of CNN models (M = 3, N = 3) with/without referring to the object mask for the *Bunny* object made of 30 different materials. Images in the second row present samples of 5 representative materials.

To test our model on different materials, we use a synthetic dataset (*Sphere* and *Bunny* object), rendered with 100 different BRDFs. We calculate the average MAE values of various CNN models with/without referring to the masked lighting

maps for 100 different materials, as illustrated in Table 2. It is observed that the performance of the per-pixel CNN model (M = 3, N = 1) remains almost unchanged after incorporating the masked lighting map. In comparison, the CNN-based models performing intraframe spatial convolutions (setting N = 3) achieved more accurate surface normal estimation results by referring to the binary object mask, which defines pixels of target and background. For instance, the average MAE of 100 different materials for Sphere is significantly reduced from 5.97° to  $4.51^{\circ}$  for the model using M = 1, N = 3 and from  $5.73^{\circ}$  to  $4.16^{\circ}$  for the model using M = 3, N = 3. The quantitative evaluation results illustrate the importance of integrating the easily obtained object mask in CNN models that involve intraframe spatial convolutions, eliminating adverse interference from invalid background regions for high-accuracy surface normal estimation.

In Fig. 5, we show the calculated MAE values using CNN models (M=3, N=3) with/without referring to the object mask for *Bunny* objects made of 30 representative materials. Note that the reflectance observations of background pixels remain zero, which presents similar variation patterns to those of object pixels with inadequate reflectance observations (e.g., made of highly absorptive dark materials). As a result, utilizing the binary object mask to eliminate adverse interference from invalid background regions leads to a significant increase in MAE values for objects of dark materials such as ss400 and specular-black-phenoli, as shown in the right part of Fig. 5. In comparison, such improvement is almost neglectable for the Bunny object made of light materials such as pink-felt and silver-paint, as illustrated in the left part of Fig. 5.



**Fig. 6.** Qualitative results of CNN models with/without referring to the object mask for real-world objects (*Bear*, *Reading* and *Pot2*) in the DiLiGenT dataset. More accurate normal estimation results are achieved for regions with cast shadows by referring to the easily obtained object mask. We purposely show error maps in the dashed boxes to better visualize normal estimation results in regions with complex geometry and obvious cast shadows.

In Fig. 6, we visualize the surface normal estimation results with/without considering the object mask for real-world objects

(*Bear*, *Reading* and *Pot2*) in the DiLiGenT dataset. More accurate surface normal estimation results are generated in regions with complex geometries and obvious cast shadows by referring to the easily obtained object mask.

# 4.2.3. Setting kernel sizes

Table 3: The comparative results (MAE) of MT-PS-CNN models using different M and N parameters on objects in the DiLiGenT benchmark.

Kerne	el Size	BALL	CAT	POT1	BEAR	POT2	BUDD.	GOBL.	READ.	COW	HARV.	Avg.
	M=1	2.97	6.23	7.90	7.13	10.24	8.23	9.30	13.37	10.01		
N=3	M=3	2.82	5.95	6.91	5.46	8.10	7.05 <b>6.85</b>	8.38	11.85	8.16	14.09	7.88
IN=3	M=5	2.29	5.87	6.92	5.79	6.89	6.85	7.88	11.94	7.48	13.71	7.56
	M=7	2.21	5.46	6.53	5.62	7.42	6.99	7.86	12.56	7.30	13.93	7.59
	N=1	2.80	5.63	6.74	6.94	9.23	7.42	10.28	12.51	12.32	15.22	8.91
M=5	N=3	2.29	5.87	6.92	5.79	6.89	6.85	7.88	11.94	7.48	13.71	7.56
	N=5	2.30	6.57	7.16	5.11	7.97	7.13	8.24	12.51	7.45	13.60	7.80

The kernel sizes of 3D convolutional layers (M and N) in inter- and intraframe feature extractors are critical parameters to determine the complexity and performance of the proposed model. The performances of MT-PS-CNN models using different M and N parameters are experimentally evaluated, and the comparative results (MAE values) on objects in the DiLi-GenT benchmark are shown in Table 3. It is observed that networks using larger kernel sizes generally produce lower MAE values. A more complex model (considering more adjacent frames and more neighbouring pixels) provides a better expressive/generalization ability to achieve more accurate surface normal estimation results. However, the improvements become insignificant when M and N are greater than 3. In our implementation, we set M = N = 3 to achieve a good balance between model complexity and good performance.

## 4.3. Comparisons with state-of-the-arts methods

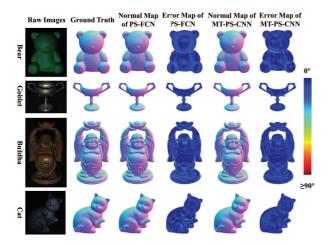
Table 4: Quantitative results of our proposed MT-PS-CNN model and state-of-the-art photometric stereo methods using the DiLiGenT benchmark dataset.

Method	BALL	CAT	POT1	BEAR	POT2	BUDD.	GOBL.	READ.	COW	HARV.	Avg.
Proposed	2.29	5.87	6.92	5.79	6.89	6.85	7.88	11.94	7.48	13.71	7.56
JU-19[32]	2.40	6.11	6.54	5.23	7.49	9.89	8.61	13.68	7.98	16.18	8.41
CH-18[12]	2.82	6.16	7.13	7.55	7.25	7.91	8.60	13.33	7.33	15.85	8.39
SI-18*[11]	2.20	4.60	5.40	12.30	6.00	7.90	7.30	12.60	7.90	13.90	8.01
TM-18[45]	1.47	5.44	6.09	5.79	7.76	10.36	11.47	11.03	6.32	22.59	8.83
HI-17[31]	2.02	6.54	7.05	6.31	7.86	12.68	11.28	15.51	8.01	16.86	9.41
ST-14[10]	1.74	6.12	6.51	6.12	8.78	10.60	10.09	13.63	13.93	25.44	10.30
IA-14[46]	3.34	6.74	6.64	7.11	8.77	10.47	9.71	14.19	13.05	25.95	10.60
BASELINE[8]	4.10	8.41	8.89	8.39	14.65	14.92	18.50	19.80	25.60	30.62	15.39

\* indicates that we use all 96 images to estimate the normal map of Bear. The result shown in SI-18 [11] (Bear 4.1) was achieved by discarding the first 20 input images.

In this subsection, we compared our proposed MT-PS-CNN model with a number of state-of-the-art photometric stereo solutions including IK-12[47], IA-14[46], ST-14[10], HI-17[31], TM-18[45], CH-18[12], SI-18[11], and JU-19[32]. The source codes or evaluation results of these methods are publicly available.

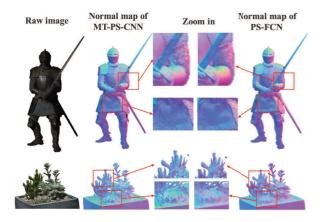
Quantitative evaluation results using the DiLiGenT benchmark dataset are provided in Table 4. When the number of input images is 96, our proposed MT-PS-CNN model achieved the lowest average MAE (7.56°) on 10 real-world objects of



**Fig. 7.** Surface normal estimation results using 96 input images on the DiLi-GenT dataset. Zoom in to assess estimation results in regions with complex geometry.

DiLiGenT. It is worth mentioning that such improvements are particularly obvious for objects with complex geometry (e.g., *Buddha*, *Reading*, and *Harvest*).

Figs. 7 and 8 show some qualitative results on two real-world DiLiGenT and Light Stage Data Gallery datasets, respectively. Our proposed MT-PS-CNN model is capable of predicting a surface normal map with rich details. Moreover, it can generate more accurate surface normal estimation results in the highlighted regions with complex geometry compared with the state-of-the-art PS-FCN method (CH-18) [12].



**Fig. 8.** Surface normal estimation results using 36 input images on the Light Stage Data Gallery dataset. The zoomed in image provides more details of the highlighted regions.

Table 5: Parameter number and running time of a number of state-of-the-art deep learning-based photometric stereo approaches.

	SI-18[11]	CH-18[12]	MT-PS-CNN
Parameter number	2.65M	2.21M	0.72M
Running time (s)	19.1	2.1	0.27

Table 5 shows the parameter number and running time of MT-PS-CNN and two representative deep learning-based pho-

tometric stereo approaches on a PC with GeForce GTX 1080Ti and 96 GB RAM. We repeat the estimation process 10 times and compute the average running time (the forward time of the network). For a fair comparison, different methods are applied for the normal estimation of the croped *BallPNG* in DiLiGenT, and the input number is 96. Following the setting in SI-18 [11], we set the number of different rotations for the rotational pseudoinvariance to 1. Our proposed MT-PS-CNN model achieves the highest normal estimation accuracy using significantly fewer parameters and runs much faster. Thus, our model is more suitable for real-time implementation in memory-restricted devices.

Another noticeable advantage of the proposed MT-PS-CNN is that it performs robustly when the number of input images significantly changes. In Table 6, we show the normal estimation results (the average MAE of 10 objects in the DiLiGenT dataset) of different deep-learning-based photometric stereo methods using 96, 16, 10, 8 and 6 input images. Note that JU-19[32] is designed to decrease the demands on the number of images for the photometric stereo task by learning the most informative images under different illumination conditions; thus, it performs the best based on 6 and 8 input images. However, its performance is suboptimal when processing dense input images. In comparison, our proposed method performs consistently well based on images captured in both sparse and dense lighting distributions. As illustrated in Table 7, MT-PS-CNN performs significantly better than other alternatives for objects with complex geometry (e.g., Buddha, Reading, and Harvest), which represents a more challenging normal estimation task.

Table 6: Surface normal estimation results using different numbers of input images (96, 16, 10, 8, 6) on the DiLiGenT dataset. We calculate the average MAE for 10 objects.

Method	96	16	10	8	6
Proposed	7.56	8.82	9.84	10.75	12.30
JU-19[32]	8.41	9.66	10.02	10.39	12.16
CH-18[12]	8.39	9.37	10.33	11.13	12.56

Table 7: Surface normal estimation results of objects in the DiLiGenT dataset using 10 illumination directions. We randomly selected 10 images from 96 inputs and calculated the average MAE over 10 trials.

Method	BALL	CAT	POT1	BEAR	POT2	BUDD.	GOBL.	READ.	COW	HARV.	Avg.
Proposed	4.20	7.30	8.78	8.59	9.85	8.25	10.44	13.17	10.84	16.97	9.84
JU-19[32]	3.97	6.70	7.30	8.73	9.74	11.36	10.46	14.37	10.19	17.33	10.02
CH-18	4.26	8.12	9.84	8.07	9.29	9.24	10.61	15.11	9.90	18.90	10.33

We also validate the effectiveness of our proposed MT-PS-CNN model on images captured using our own photometric stereo data acquisition system. Fig. 9 shows the comparative performance of PS-FCN [12], CNN-PS [11] and the proposed MT-PS-CNN. It is observed that the PS-FCN model utilizes intraframe information to effectively suppress interference of specular highlights. However, some important details (e.g., scratches on the steel surface) are adversely smoothed out, as shown in Fig. 9. In comparison, the CNN-PS model generates

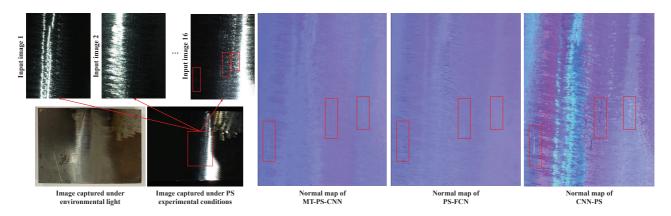


Fig. 9. Normal map comparison of different methods on polished steel surfaces.

a normal map with rich textures/details, but its performance is significantly affected by specular highlights due to the lack of consideration of spatial information. Our proposed MT-PS-CNN model utilizes both inter- and intraframe information and thus can combine the advantages of the above two methods.

#### 5. Conclusion

In this paper, we present a complete photometric stereo acquisition/processing framework including a multilight source illumination and acquisition system and a two-stage CNN architecture to extract inter- and intraframe feature representations for high-quality surface normal estimation of non-Lambertian objects. We experimentally investigate a number of network design alternatives to identify the optimal scheme (T-IRFE-IAFE) to deploy interframe and intraframe feature extraction modules for the photometric stereo problem. Moreover, we integrate the easily obtained object mask in intraframe spatial convolutions to improve the accuracy of normal estimation for surfaces made of highly absorptive dark materials or with obvious cast shadows. The advantages of the proposed MT-PS-CNN model include producing more accurate normal estimation results using significantly few parameters and performing robustly under both dense and sparse image capturing configurations. The source code of the MT-PS-CNN model will be made publicly available.

We experimentally observed that the order of input images will affect the performance of our method based on inter- and intraframe 3D convolution. In the future, we plan to systematically investigate the optimal strategy to process image sequences to achieve high-accuracy and robust surface normal estimation. Moreover, our proposed MT-PS-CNN is a lightweight model, and it is possible to stack more convolutional layers/modules or adopt more complex network architectures (e.g., PS-FCN<sup>+N</sup>[48], PX-NET[49]) to further improve its accuracy.

# References

[1] Zhao Song, Ying Nie, and Zhan Song. Photometric stereo with quasipoint light source. *Optics and Lasers in Engineering*, 111:172–182, 2018.

- [2] Jesús Villa and Juan B Hurtado-Ramos. Surface shape estimation from photometric images. *Optics and lasers in engineering*, 42(4):461–468, 2004.
- [3] Y. Fang, G. Ding, W. Wen, F. Yuan, Y. Yang, Z. Fang, and W. Lin. Salient object detection by spatiotemporal and semantic features in realtime video processing systems. *IEEE Transactions on Industrial Electronics*, 67(11):9893–9903, 2020.
- [4] Long Ma, Jirui Liu, Xin Pei, Yanmin Hu, and Fengming Sun. Calibration of position and orientation for point light source synchronously with single image in photometric stereo. *Optics express*, 27(4):4024–4033, 2019.
- [5] Limin Xie, Zhan Song, Guohua Jiao, Xinhan Huang, and Kui Jia. A practical means for calibrating an led-based photometric stereo system. *Optics and Lasers in Engineering*, 64:42–50, 2015.
- [6] Haowen Zhou, Xiaomeng Sui, Liangcai Cao, and Partha P Banerjee. Digital correlation of computer-generated holograms for 3d face recognition. Applied optics, 58(34):G177–G186, 2019.
- [7] Pei Zhou, Jiangping Zhu, and Zhisheng You. 3-d face registration solution with speckle encoding based spatial-temporal logical correlation algorithm. *Optics express*, 27(15):21004–21019, 2019.
- [8] Robert J. Woodham. Photometric Method For Determining Surface Orientation From Multiple Images. *Optical Engineering*, 19(1):139–144, feb 1980
- [9] D.B. Goldman, Brian Curless, Aaron Hertzmann, and S.M. Seitz. Shape and spatially-varying BRDFs from photometric stereo. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume I, pages 341–348 Vol. 1. IEEE, 2005.
- [10] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. Bi-Polynomial Modeling of Low-Frequency Reflectances. *IEEE Transac*tions on Pattern Analysis and Machine Intelligence, 36(6):1078–1091, jun 2014.
- [11] Satoshi Ikehata. CNN-PS: CNN-based Photometric Stereo for General Non-Convex Surfaces. In Proceedings of the European conference on computer vision (ECCV), pages 614–629, 2018.
- [12] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. PS-FCN: A Flexible Learning Framework for Photometric Stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19. jul 2018.
- [13] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan Yee K.K. Wong. Self-calibrating deep photometric stereo networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8731–8739, 2019.
- [14] Jens Ackermann and Michael Goesele. A Survey of Photometric Stereo Techniques. *Foundations and Trends*® *in Computer Graphics and Vision*, 9(3-4):149–254, 2015.
- [15] Fredric Solomon and Katsushi Ikeuchi. Extracting the shape and roughness of specular lobe objects using four light photometric stereo. In Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, number 7597, pages 466–471. IEEE Comput. Soc. Press.
- [16] E. North Coleman and Ramesh Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics and Image Processing*, 18(4):309–328, apr 1982.

- [17] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma. Robust Photometric Stereo via Low-Rank Matrix Completion and Recovery. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 703–717. 2011.
- [18] Yasuhiro Mukaigawa, Yasunori Ishii, and Takeshi Shakunaga. Analysis of photometric factors based on photometric linearization. JOSA A, 24(10):3326–3334, 2007.
- [19] Daisuke Miyazaki, Kenji Hara, and Katsushi Ikeuchi. Median Photometric Stereo as Applied to the Segonko Tumulus and Museum Objects. International Journal of Computer Vision, 86(2-3):229–242, jan 2010.
- [20] K. E. Torrance and E. M. Sparrow. Theory for Off-Specular Reflection From Roughened Surfaces\*. *Journal of the Optical Society of America*, 57(9):1105, sep 1967.
- [21] Greg Kay and Terry Caelli. Estimating the parameters of an illumination model using photometric stereo, 1995.
- [22] Georghiades. Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo. In *Proceedings Ninth IEEE International Conference on Computer Vision*, volume 2, pages 816–823 vol.2. IEEE, 2003.
- [23] Robert L. Cook and Kenneth E. Torrance. A reflectance model for computer graphics. In *Proceedings of the 8th annual conference on Computer graphics and interactive techniques SIGGRAPH '81*, volume 15, pages 307–316, New York, New York, USA, 1981. ACM Press.
- [24] Hin-Shun Chung and Jiaya Jia. Efficient photometric stereo on glossy surfaces with wide specular lobes. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, jun 2008.
- [25] Shape and Spatially-Varying BRDFs from Photometric Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(6):1060– 1071, jun 2010.
- [26] Bui Tuong Phong. Illumination for computer generated pictures. Communications of the ACM, 18(6):311–317, jun 1975.
- [27] James F. Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques SIGGRAPH '77*, pages 192–198, New York, New York, USA, 1977. ACM Press.
- [28] Addy Ngan, Frédo Durand, and Wojciech Matusik. Experimental analysis of brdf models. *Rendering Techniques*, 2005(16th):2, 2005.
- [29] Michael M Stark, James Arvo, and Brian Smits. Barycentric parameterizations for isotropic brdfs. *IEEE transactions on visualization and computer graphics*, 11(2):126–138, 2005.
- [30] Aaron Hertzmann and Steven M Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Trans*actions on Pattern Analysis and Machine Intelligence, 27(8):1254–1264, 2005
- [31] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep Photometric Stereo Network. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pages 501–509. IEEE, 2017.
- [32] Junxuan Li, Antonio Robles-Kelly, Shaodi You, and Yasuyuki Matsushita. Learning to Minify Photometric Stereo. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), volume 2019-June, pages 7560–7568. IEEE, jun 2019.
- [33] Qian Zheng, Yiming Jia, Boxin Shi, Xudong Jiang, Ling-Yu Duan, and Alex Kot. SPLINE-Net: Sparse Photometric Stereo Through Lighting Interpolation and Normal Estimation Networks. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8548–8557. IEEE, oct 2019.
- [34] Xi Wang, Zhenxiong Jian, and Mingjun Ren. Non-Lambertian Photometric Stereo Network Based on Inverse Reflectance Model With Collocated Light. *IEEE Transactions on Image Processing*, 29(5):6032–6042, 2020.
- [35] Yasuhiro Mukaigawa, Yasunori Ishii, and Takeshi Shakunaga. Analysis of photometric factors based on photometric linearization. *Journal of the Optical Society of America A*, 24(10):3326, oct 2007.
- [36] Junxuan Li, Antonio Robles-Kelly, Shaodi You, and Yasuyuki Matsushita. Learning to Minify Photometric Stereo. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), volume 2019-June, pages 7560–7568. IEEE, jun 2019.
- [37] Xi Wang, ZhenXiong Jian, and Mingjun Ren. Non-Lambertian Photometric Stereo Network Based on Inverse Reflectance Model With Collocated Light. *IEEE Transactions on Image Processing*, 29(5):6032–6042, 2020.

- [38] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1586–1594, 2017
- [39] Olivia Wiles and Andrew Zisserman. Silnet: Single-and multi-view reconstruction by learning from silhouettes. In BMVC, 2017.
- [40] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K Wong, and Yasuyuki Matsushita. What is learned in deep uncalibrated photometric stereo? In European Conference on Computer Vision, pages 745–762. Springer, 2020.
- [41] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. ACM Transactions on Graphics, 22(3):759, jul 2003.
- [42] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE Conference* on Computer Vision and Pattern Recognition, pages 3707–3716, 2016.
- [43] Charles-Félix Chabert, Per Einarsson, Andrew Jones, Bruce Lamond, Wan-Chun Ma, Sebastian Sylwan, Tim Hawkins, and Paul Debevec. Relighting human locomotion with flowed reflectance fields. In ACM SIG-GRAPH 2006 Sketches, pages 76–es. 2006.
- [44] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6450–6459. IEEE, jun 2018.
- [45] Tatsunori Taniai and Takanori Maehara. Neural Inverse Rendering for General Reflectance Photometric Stereo. 35th International Conference on Machine Learning, ICML 2018, 11:7731–7740, feb 2018.
- [46] Satoshi Ikehata and Kiyoharu Aizawa. Photometric Stereo Using Constrained Bivariate Regression for General Isotropic Surfaces. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 2187–2194. IEEE, jun 2014.
- [47] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa. Robust photometric stereo using sparse regression. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, number 2, pages 318–325. IEEE, jun 2012.
- [48] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee Kenneth Wong. Deep photometric stereo for non-lambertian surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.
- [49] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla. Px-net: Simple, efficient pixel-wise training of photometric stereo networks. arXiv preprint arXiv:2008.04933, 2020.