SpectraM-PS: Spectrally Multiplexed Photometric Stereo under Unknown Spectral Composition

Satoshi Ikehata[1,2] and Yuta Asano[1]

[1] National Institute of Informatics, Tokyo, Japan
[2] Tokyo Institute of Technology, Tokyo, Japan

Abstract. In this paper, we present a groundbreaking spectrally multiplexed photometric stereo approach for recovering surface normals of dynamic surfaces without the need for calibrated lighting or sensors, a notable advancement in the field traditionally hindered by stringent prerequisites and spectral ambiguity. By embracing spectral ambiguity as an advantage, our technique enables the generation of training data without specialized multispectral rendering frameworks. We introduce a unique, physics-free network architecture, SpectraM-PS, that effectively processes multiplexed images to determine surface normals across a wide range of conditions and material types, without relying on specific physically-based knowledge. Additionally, we establish the first benchmark dataset, SpectraM14, for spectrally multiplexed photometric stereo, facilitating comprehensive evaluations against existing calibrated methods. Our contributions significantly enhance the capabilities for dynamic surface recovery, particularly in uncalibrated setups, marking a pivotal step forward in the application of photometric stereo across various domains.

Keywords: Spectrally Multiplexed Photometric Stereo · Dynamic Surface Recovery · Multispectral Photometric Stereo

1 Introduction

Recovering detailed normals of dynamic surfaces is essential for monitoring various processes: in manufacturing, it helps in tracking wear and tear of machine parts; in agriculture, it allows for the observation of crop growth through changes in leaf geometry; and in sports engineering, it aids in improving equipment design and safety by analyzing how surfaces deform upon impact.

Photometric Stereo (PS) [57,64] derives object surface normals from observations under different lighting conditions at a fixed viewpoint. Despite decades of progress, the requirement for objects to stay stationary during lighting changes challenges the recovery of dynamic surfaces, essential for analyzing temporal surface deformations. PS researches have employed *spectral multiplexing* for dynamic

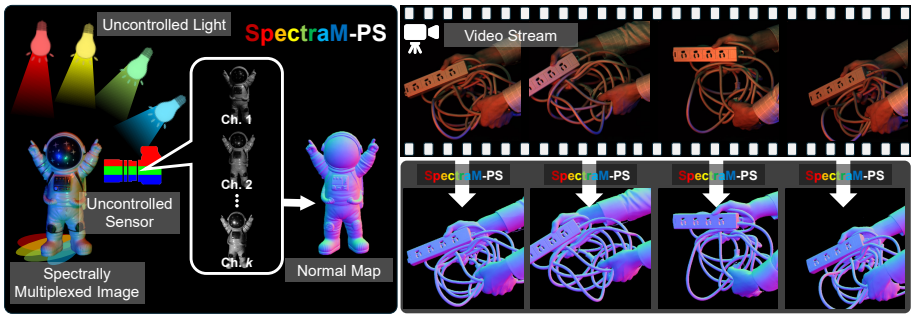

Fig. 1: (Left) Illustration of our SpectraM-PS. Our method recovers a surface normal map from a spectrally multiplexed image. The spectral/spatial composition for generating the observations is unknown. There is potential for a mismatch between the sensor's spectral sensitivity and the light source's spectral distribution, which may lead to crosstalk. (Right) By applying our method to individual frames of a video, the normal map of dynamic surfaces can be recovered.

surface recovery [10,20,32,37,51,62]—a technique originally used in telecommunications and spectroscopy [31]. This technique utilizes the varying wavelengths of light to multiplex and subsequently demultiplex signals within a single sensor, thereby increasing the capacity for information transmission.

Historically, spectrally multiplexed photometric stereo is often referred to as color photometric stereo [7,14,20,32,33], specifically when objects are illuminated with monochromatic red, green, and blue lights from *various angles*, captured in the camera's RGB channels. Each channel is then treated as an observation under a distinct lighting for PS analysis. This technique has been further extended to not only RGB but also any number of spectral bands and is specifically referred to as multispectral photometric stereo [17,18,46]. These techniques enable dynamic surface recovery by processing each temporal multi-channel frame separately.

Despite their potential in dynamic surface recovery, current spectrally multiplexed photometric stereo methods face stringent prerequisites that limit their practicality. These include the necessity for precisely calibrated directional lighting in controlled environments [7, 14, 17, 20, 46] and sensors with aligned spectral sensitivities [32, 33]. Furthermore, they make strong assumptions about the surface, requiring it to be convex, integrable, Lambertian, and exhibit uniform chromaticity [20,37]. By contrast, recent PS methods without spectral multiplexing support non-Lambertian surfaces [29,55], spatially-varying materials [12,23], and the use of uncalibrated lighting [11,25,26,54]. This disparity arises from the challenge where identical observations are produced by different spectral compositions of light, surface and sensor [22, 50], a phenomenon absent in conventional PS due to constant spectral compositions of them across images. Recently, Guo *et al*. [17,18] thoroughly explored how the spectral ambiguity renders spectrally multiplexed photometric stereo ill-posed, necessitating severely unpractical conditions on light, surface, and sensor to resolve the ambiguity.

In this work, we propose a spectrally multiplexed photometric stereo method that recovers normals directly from multiplexed observations produced by *un-*

known composition of lights, surface, and sensor (See Fig. 1-left), drawing inspiration from recent data-driven photometric stereo methods [25, 26]. While prior works [18, 46] have considered the spectral ambiguity harmful and something that must be resolved, we demonstrate that it can even be beneficial for a data-driven approach as it compacts the input space and allows for the generation of training data without a multispectral rendering framework. Trained on spectrally composed observations, our generic, physics-free architecture directly maps a single multiplexed image with an order-agnostic, arbitrary number of channels to object surface normals without the need for calibrating lights and sensors, and without imposing severe constraints on surface reflectance and geometry. By applying our method to individual frames of a video, dynamic surface recovery via spectrally multiplexed photometric stereo in uncalibrated, uncontrolled scenarios is achieved as illustrated in Fig. 1 (right).

While numerous benchmarks exist for conventional photometric stereo [52,56, 63], not a single benchmark is available for spectrally multiplexed PS. Therefore, we have created the first real benchmark dataset, namely *SpectraM*, for this task. For comparative evaluations with calibrated methods such as [18, 46], we carefully calibrated directional light sources of different wavelengths, including their directions. We implemented five different difficulty settings by varying the type of light sources (RGB vs NIR) and whether individual light sources were illuminated independently or simultaneously, catering to both ideal conditions without channel crosstalk and more realistic conditions with channel crosstalk.

Our contributions are summarized as follows: (1) We pioneer the use of spectrally multiplexed photometric stereo for recovering dynamic surfaces in uncalibrated setups, employing a data-driven approach to overcome spectral ambiguity, a significant barrier in prior work. (2) We introduce a unique, physics-free neural network, *SpectraM-PS* (**Spectra**lly **M**ultiplexed **PS**), that recovers surface normals from a spectrally multiplexed image, capable of handling images with any number of order-agnostic channels. (3) We demonstrate how spectral ambiguity restricts the input space for training data generation, offering a strategy for efficient dataset creation without the need for multispectral rendering. (4) We create the first evaluation benchmark, *SpectraM*, for this domain, showing our method's superiority over current calibrated spectrally multiplexed photometric stereo techniques.

2 Related Work

Temporally Multiplexed Photometric Stereo (Conventional). From a communication perspective, conventional photometric stereo, as originally proposed by Woodham [64], employs a *time multiplexing* strategy to recover static surfaces. This method involves temporally varying lighting conditions while capturing images from a fixed viewpoint. Since the same light sources and sensor always provide observations, image differences stem solely from changes in light direction and intensity. This approach simplifies addressing complex conditions such as cast shadows [27, 28], non-Lambertian surfaces [16, 29], non-convex sur-

faces [23], and uncalibrated lighting [11, 36, 54]. Recently, learning-based methods [11–13, 23–25, 34, 36, 39–42, 44, 53, 58, 59, 68, 69] have emerged as an effective alternative, addressing challenges faced by traditional, physics-based approaches [6, 8, 16, 19, 21, 28, 45, 65]. These data-driven methods regress normal maps from observations utilizing techniques such as observation map regression [23, 30], set pooling [12, 13], graph neural networks [68], Transformer [24], and neural rendering for inverse rendering optimization [40, 41, 58]. Notably, the introduction of universal photometric stereo methods [25, 26] has enabled the handling of unknown, spatially-varying lighting in a purely data-driven framework. Inspired by these advancements, our work aims to regress normals from observations under unknown light, surface, and sensor conditions.

Spectrally Multiplexed Photometric Stereo. Despite its potential for dynamic surface recovery, spectrally multiplexed photometric stereo [15, 37] has remained less explored than its mainstream counterpart, primarily due to notable limitations.

Lighting Constraints: Existing methods necessitate multiple directional lights in controlled settings, contrasting the flexibility of temporally multiplexed techniques that adapt to diverse lighting conditions [26, 47, 49]. They typically require pre-calibrated light directions and distinct light source spectra to prevent channel crosstalk. In contrast, our approach accommodates uncontrolled lighting scenarios without the need for predefined or calibrated setups.

Surface and Sensor Constraints: Prior works assume significant limitations on surfaces, such as Lambertian, convex, and uniform properties [15, 20, 37]. Recent advances like Lv *et al*. [46] extend to non-Lambertian surfaces but still require uniform materials. Sensor requirements typically involve narrow-band spectral responses and a fixed number of channels, limiting flexibility. Our approach leverages a data-driven model, training neural networks on synthetic data to handle complex surfaces and varied spectral sensor responses.

Data-driven Methods: To our knowledge, there are few data-driven methods for this task [32, 33, 46]. Previous studies, such as those by Ju *et al*. [32, 33], require identical spectral and spatial lighting conditions during both training and testing, greatly restricting their practicality. ELIE-Net [46] permits variability in training and test setups; however, strong assumptions on both light and surfaces prohibitively limit its applications. Our model, on the other hand, eschews explicit lighting models in favor of learning direct input-output relationships, allowing for accurate predictions under varied and unknown spectral compositions and supporting materials with spatially diverse properties. Furthermore, unlike ELIE-Net's reliance on spectral BRDF datasets, our training approach utilizes assets akin to those employed in conventional photometric stereo.

3 Problem Statement

Given a single image $I \in \mathbb{R}^{h \times w \times k}$ captured by a static k-channel orthographic sensor, along with an optional object mask $M \in \mathbb{R}^{h \times w}$, the objective of spectrally multiplexed photometric stereo is to recover the surface normals of the

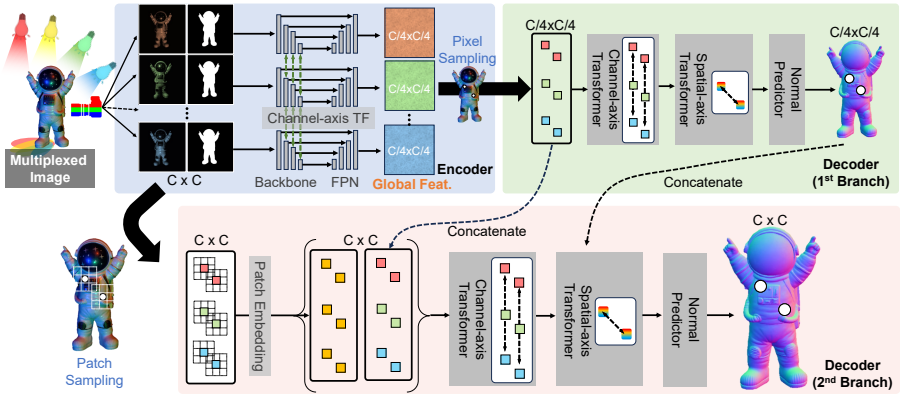

Fig. 2: SpectraM-PS involves decomposing a spectrally multiplexed image into independent channels. The Global Feature Encoder extracts a feature map from each channel. The surface vector is then recovered by the Dual-scale Surface Normal Decoder at each pixel. We adopt a dual-scale approach to preserve the entire shape, while employing patch-embedding techniques to enhance local surface details.

object, $N \in \mathbb{R}^{h \times w \times 3}$[3]. The object is supposed to be illuminated by multiple light sources, each with unique spatial and spectral properties. Previous studies have typically assumed an equal number of light sources and sensor channels, with each light source's wavelength precisely matching the spectral response of a single channel, thereby precluding any channel crosstalk, and with the directions of lights predetermined. In contrast, we do not presuppose the spatial distribution of illumination nor require the spectrum of each light source to be exclusively aligned with the spectral responses of the sensor channels, thus permitting channel crosstalk. This distinction is elaborated in subsequent sections.

4 Method

We propose and tackle the problem of spectrally multiplexed photometric stereo from a single image with multiple channels, produced by an unknown spectral/spatial composition of the sensor, light, and surface. To build such a method, we train neural networks to directly infer the normal map from an image.

Our method addresses two challenges: (1) a physics-free architecture that accepts a varying number of spectral channels and is agnostic to their order, and (2) an effective approximation of the spectrally multiplexed image for efficient training. We consider (2) to be of significant importance, yet it remains largely unexplored. Synthesizing spectrally multiplexed images in a physically accurate manner is prohibitively challenging, owing to the increased complexity of their parameter spaces and the scarcity of 3D assets with detailed spectral properties,

[3] It should be noted that unlike conventional PS, reflectance recovery generally falls outside the scope of spectrally multiplexed PS due to its inherently ill-posed nature.

as well as the complex nature of light-surface interactions across different wavelengths. To address this issue, developing an efficient approximation method for rendering spectrally multiplexed images using common RGB image rendering techniques is crucial.

4.1 Physics-free Spectrally Multiplexed PS Network (SpectraM-PS)

The architecture of SpectraM-PS is illustrated in Fig. 2. Drawing inspiration from established Transformer-based photometric stereo networks [24–26], we integrate an encoder to first extract the global features and a decoder to estimate per-pixel surface normals. The architecture derives normals solely from the input image and mask, without prior light information. This indicates that the architecture focuses the network's learning objective on the relationship between input and output without relying on physics-based principles, unlike prior works.

In our model, all the interactions among features from different sensor channels are employed by näive Transformer [61] in similar to [24–26]. Transformer functions by mapping input features to query, key, and value vectors of equal dimensions. These vectors are processed through a multi-head self-attention mechanism, utilizing a softmax layer, followed by a feed-forward network comprising two linear layers. Both the input and output layers maintain identical dimensionality, with the inner layer having twice the dimension of the input. Each layer is surrounded by a residual connection, succeeded by layer normalization [67]. The advantage of employing Transformers in photometric stereo networks lies in their capability to facilitate complex interactions among intermediate features, a task unachievable with simple operations like pooling [11–13, 46] and observation map [23, 30]. Additionally, the token-based attention mechanism allows for different number of input tokens (*i.e.*, sensor channels) between training and test phases and ensures that the results are independent on the order of tokens.

Building on the established Transformer-based architecture [26] for temporally multiplexed photometric stereo, we extend its scope to a spectrally multiplexed one. To accommodate a variable number of channels and eliminate dependency on their order, an input spectrally multiplexed image is first *split* into individual channels, each of which is concatenated with an object mask (If no mask is provided, replace with a matrix of ones.) and then input into the same encoder of a neural network. This approach is distinctly different from traditional methods that encode an input image as it is in neural networks [32, 33]. Then, at **Preprocessing**, we normalize each channel by dividing it by a random value between its maximum and mean. Each channel and mask are resized or cropped to a resolution $(c \times c)$ that is a multiple of 32 to be input into the multi-scale encoder. **Global Feature Encoder** first applies a backbone network (*i.e.*, ConvNeXt-T [43]) to individualy encode the concatenation of each channel and mask, then uses Transformer layers for channel-axis (*i.e.*, sensor channel) feature communication across scales (the number of Transformer layers is {0, 1, 2, 4} at {1/4, 1/8, 1/16, 1/32} scales, hidden dimensions are same with input dimensions), and finally, a feature pyramid network [66] for integrating features

at different levels. Note that the design of encoder is almost the same as [25,26], except that images are replaced by sensor channels, so details are omitted.

Given global features $\in \mathcal{R}^{k\times c/4\times c/4\times 256}$, our novel **Dual-scale Surface Normal Decoder** adopts a dual-scale strategy for predicting point-wise surface normals at m (*i.e.*, 2048) sampled locations at the original resolution within the object mask. The first branch recovers low-frequency surface normals at the feature map resolution ($\frac{c}{4} \times \frac{c}{4}$). Concretely, all global features corresponding to each sample location are processed by five channel-axis Transformer layers (with a 256 hidden dimension) and are pooled via Pooling-by-Multihead-Attention (PMA) [38] using an additional channel-axis Transformer layer (with a 384 hidden dimension). To enhance spatial communication, two *spatial*-axis Transformer layers (with a 384 hidden dimension) inspired by Ikehata [26] are employed (*i.e.*, Transformer is employed among samples at different locations), with a final MLP (384→192→3) predicting the low-frequency normals at sampled locations. The second branch focuses on high-resolution normal recovery, using patch embedding for local context at the same m locations, with $w \times w$ patches ($w = 21$) processed by an MLP (with a 256 hidden dimension) and two layer norms. These patches, concatenated with bilinearly interpolated global features, pass through five channel-axis Transformer blocks (with a 256 hidden dimension), PMA (with a 384 hidden dimension), and are merged with the first branch output normals into 387-dimensional vectors. Two additional spatial-axis Transformer layers (with a 384 hidden dimension) enable non-local interactions, culminating in a final MLP (384→192→3) for high-resolution normals, normalized to unit vectors. The complete normal map is formed by merging all the vectors from different sample sets.

It should be noted that while SDM-UniPS [26] targets temporally multiplexed PS with tens of images, and its decoder performs normal estimation purely on a pixel basis. In contrast, spectrally multiplexed PS deals with fewer channels (*e.g.*, three with RGB sensors), making a pixel-basis architecture less effective. Therefore, we use *patch embedding* at the patch-basis decoder to capture fine details with a *dual-scale architecture* for preserving overall shape. Without a dual-scale design, the recovery of surface normals becomes overly influenced by local image textures captured through patch embedding. This leads to a failure in preserving the entire shape, resulting in a significant reduction in accuracy. Our motivation is supported by Fig. 3 (left), where SDM-UniPS [26] fails to recover fine details with six temporally multiplexed images, while our architecture produces a more plausible normal map.

4.2 Efficient Training Strategy Utilizing Spectral Ambiguity

Aligning the training and test data domains in neural networks is essential for optimal model performance [9,60]. However, rendering spectrally multiplexed data poses challenges due to the scarcity of multispectral Bidirectional Reflectance Distribution Functions (BRDFs). In reality, ELIE-Net [46] was trained using only 51 measured isotropic spectral BRDFs. On the other hand, given the availability of various large isotropic BRDF databases [1–3,5,48], we seek to explore

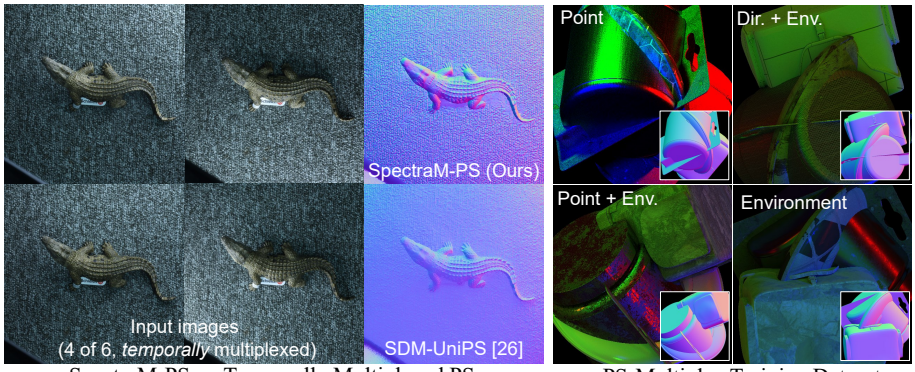

Fig. 3: (Left) Comparison of SpectraM-PS and SDM-UniPS [26] on six *temporally* multiplexed PS images. Due to the patch-wise basis of SpectraM-PS, fine details are better recovered. (Right) Illustration of different lighting conditions in PS-Multiplex.

utilizing these datasets for training our model, leveraging the fact that our network does not distinguish images based on their physically-based principles. In this section, we highlight how RGB images serve as a practical approximation, simplifying the complexity inherent in multispectral imaging.

We begin the discussion by characterizing multispectral imaging. Assuming that the surface doesn't emit light and only reflections on surface are considered, the image formation model is described as follows [35]:

$$I_{(s,p)} = \int_{\Omega} (\omega_i^{\mathsf{T}} n_p) \int_0^{\infty} S_s(\lambda) f_p(\omega_i, \omega_o, \lambda) L_p(\omega_i, \lambda) \mathrm{d}\lambda \mathrm{d}\omega_i. \tag{1}$$

In this equation, $I_{(s,p)}$ denotes the incoming spectral radiance at the sensor s (or s-th channel) from a surface point p. The term f_p represents BRDF, L_p the incident light intensity at the surface point, and λ the wavelength of the incident light. The symbols ω_i and ω_o denote the directions of incident and reflected light, respectively. $S_s(\lambda)$ refers to the spectral sensitivity of the sensor s at wavelength λ, n_p is the surface normal, and Ω represents the hemisphere over which incident light directions are possible. The integral sums over all incident directions and wavelengths. It is important to note that the incident light intensity L_p depends not only on the direct contribution from light sources but also on the visibility of light (e.g., attached and cast shadows) and indirect illuminations.

Eq. (1) illustrates the concept of spectral ambiguity, showing that an infinite number of combinations of $S_s(\lambda)$, $f_p(\omega_i, \omega_o, \lambda)$, and $L_p(\omega_i, \lambda)$ can result in the same spectral radiance, including narrowband compositions. In other words, with spectral ambiguity, a single observation $I_{(s,p)}$ can encompass the observations for all spectral compositions that satisfy the equation (*i.e.*, metamerism [22, 50]). This perspective justifies the theory of substituting multispectral images, which possess a broad parameter space, with narrowband RGB images. It is worth mentioning that channel crosstalk primarily affects the incident light intensity, consequently distorting the product of $S_s(\lambda) \cdot f_p(\omega_i, \omega_o, \lambda) \cdot L_p(\omega_i, \lambda)$ in Eq. (1).

This implies that observations influenced by spectral crosstalk can still be equivalently represented using a narrowband setup under spectral ambiguity. In the experiments, we demonstrate that our model, trained on three narrowband observations can be applied to multiplexed data with channel crosstalk. To realize this approximation, we rendered a large number of three-channel narrowband images using the path-tracing algorithm in Blender [4], where up to 10-bounce reflections are permitted, based on common 3D assets [2] for RGB rendering. Following the rendering pipeline described in [26], we rendered objects by combining three different lighting models: directional, point, and environmental (five combinatorial settings in total as shown in Fig. 3). To simulate spectrally multiplexed images, we defined R, G, and B light sources and illuminated the surface in a multiplexed manner. It is important to note that the rendered RGB images are decomposed into three grayscale images, each of which was independently fed into the network; therefore, any wavelength-dependent information is masked. For material diversity, we adopted the method from [26], categorizing 897 AdobeStock texture maps into three groups: 421 diffuse, 219 specular, and 257 metallic textures. Four objects from a set of 410 3D AdobeStock models were randomly selected and textured with these materials. This structured approach led to the rendering of 106,374 multiplexed images along with their ground truth surface normal maps, forming the 'PS-Multiplex' dataset.

5 SpectraM14 Benchmark Dataset

Due to the lack of a benchmark for spectrally multiplexed PS, the first comprehensive evaluation dataset, named SpectraM14, is created. This dataset includes 14 objects, each exhibiting a range of optical properties such as monochromatic or multicolored appearances and diffuse or specular reflections, as depicted in Fig. 4. Our benchmark encompasses tasks under five distinct conditions, as described later.

Imaging Setup. To acquire our dataset, we utilized a color camera (FLIR GS3-U3-123S6C-C) and an NIR camera (FLIR GS3-U3-41C6NIR-C), both equipped with a 50mm lens. For the NIR camera, we used narrowband filters with wavelengths of 750nm, 850nm, 880nm, 905nm, and 940nm, and the acquired images were manually merged. Objects were placed 0.8m from the camera to approximate orthographic projection. Following conventional PS benchmarks [52,56,63], data capture occurred in a controlled, dark environment with the scene draped in black cloth to mitigate interreflection. The camera's ISO sensitivity was minimized to enhance image quality. The imaging area was further isolated using low-reflectance cloths to suppress inter-reflection. For each illumination condition, we collected six images under varying exposures to produce HDR input images. For the evaluations throughout this paper, the images are cropped using an object mask and resized to 512px $\times$ 512px.

Lighting. Six LED and three halogen light sources, positioned roughly 1 meter from the object, provided illumination. We used the "Weeylite S05 RGB Pocket Lamp" and the "NPI PIS-UHX-AIR" for lighting. This setup enabled the use of

Fig. 4: Objects in SpectraM14.

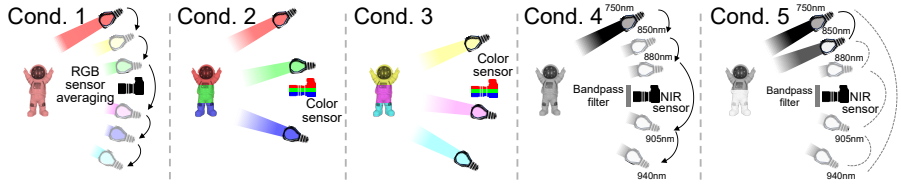

Fig. 5: Illustration of six conditions in SpectraM14.

red, green, blue, yellow, magenta, cyan, and NIR lighting, with spectra validated using a Hamamatsu Photonics Multichannel Analyzer C10027-01.

Calibration and Ground Truth Data. We measured the directions of lights using specular reflections from a mirror sphere. Light intensity was standardized across the visible spectrum by averaging RGB values from reflected light on a white target. The ground truth normals were captured with a SHINING 3D EinScan-SE scanner.

Evaluation Procedure. The design philosophy of this benchmark is to assess the robustness and adaptability of spectrally multiplexed PS methods under realistic lighting conditions, accounting for variations in channel numbers and the presence of spectral crosstalk. For a comprehensive evaluation, we designed tasks under five distinct conditions as shown in Fig. 5: **Condition 1**: Color sensor, no crosstalk condition: Six colors of light (red, green, blue, cyan, yellow, magenta) were each independently illuminated and observed with an RGB sensor. Afterward, the channels of RGB were averaged. **Condition 2**: Color sensor, weak crosstalk condition: Three colors of light (red, green, blue) were simultaneously illuminated and observed through each channel of the RGB sensor. **Condition 3**: Color sensor, strong crosstalk condition: Three colors of light (cyan, yellow, magenta) were simultaneously illuminated and observed through each channel of the RGB sensor. **Condition 4**: NIR sensor, no crosstalk condition: Light at wavelengths of 750 nm, 850 nm, 880 nm, 905 nm, and 940 nm were each independently illuminated and observed with a monochrome sensor corresponding to

each wavelength. **Condition 5**: NIR sensor, spatially-varying lighting condition: New images were created by averaging two images taken under the conditions mentioned above. The combinations were (750 nm, 850 nm), (850 nm, 880 nm), (880 nm, 905 nm), (905 nm, 940 nm), and (940 nm, 750 nm).

6 Experiment

In this section, we evaluate our method on our SpectraM14. Our method is compared with one SOTA optimization-based method [17] and one SOTA learning-based method [46]. The former introduces a closed-form solution for spectrally multiplexed photometric stereo applied to monochromatic surfaces with spatially varying (SV) albedo. The latter presents a Spectral Reflectance Decomposition (SRD) model, which disentangles spectral reflectance into geometric and spectral components for surface normal recovery under non-Lambertian spectral reflectance conditions. Unlike the compared methods, our approach does not assume a specific lighting setup, whereas both methods presume the presence of calibrated single directional light sources.

Training details. SpectraM-PS was trained from scratch on the PS-Multiplex dataset until convergence using the AdamW optimizer, with a step decay learning rate schedule that reduced the learning rate by a factor of 0.8 every ten epochs. We applied learning rate warmup during the first epoch and used a batch size of 16, an initial learning rate of 0.0001, and a weight decay of 0.05. Each batch consisted of three input training multiplexed images with three channels each. The training loss was computed using the Mean Squared Error (MSE) loss function to measure ℓ_2 errors between the predicted surface normal vectors and the ground truth surface normal vectors. We measured the reconstruction accuracy of our method by computing the mean angular errors (MAE) between the predicted and true surface normal maps, expressed in degrees.

Computational Cost. The inference time of PS methods varies with the number of pixels and channels in the input image. For Condition 2 and 3 with a $512 \times 512 \times 3$ image, the mean and standard deviation of inference times (in sec) over 14 objects in SpectraM14 benchmark were: our method (3.42/0.85), Lv *et al.* [46] (0.46/0.24) and Guo *et al.* [17] (2.38/1.10). Our architecture leads to higher computational costs; however, none of the methods were suitable for real-time processing (*e.g.*, 15 fps requires 0.06 sec/frame).

Ablation Study. We firstly validate the individual technical contributions of our training dataset (*i.e.*, PS-Multiplex) and the physics-free architecture (*i.e.*, SpectraM-PS) using a synthetic evaluation dataset. Firstly, we validate the efficacy of our training dataset, PS-Multiplex, by adapting an existing universal photometric stereo architecture designed for the conventional task (*i.e.*, SDM-UniPS [26]) to the spectrally multiplexed photometric stereo task. Since both ours and SDM-UniPS take multiple observations and an object mask as input, this adaptation straightforwardly involves training the model on PS-Multiplex by treating each channel of an image as an individual image. Subsequently, we

Table 1: Ablation analysis of the contributions of SpectraM-PS and PS-Multiplex.

Method	MAE (Uniform)		MAE (Piece-wise uniform)		MAE (Non-uniform)	
	Non-Lambertian	Lambertian	Non-Lambertian	Lambertian	Non-Lambertian	Lambertian
SDM-UniPS [26]	12.9 (4.5)	12.4 (4.7)	15.0 (5.0)	21.7 (7.0)	14.4 (5.1)	15.2 (5.9)
[26] trained on PS-Multiplex	11.1 (3.6)	11.0 (3.9)	10.5 (2.9)	12.3 (3.9)	10.6 (3.9)	11.2 (3.7)
SpectraM-PS (Ours)	**8.0** (2.7)	**8.4** (2.7)	**7.9** (2.4)	**8.9** (2.9)	**8.2** (3.2)	**8.0** (2.5)

compare this model against our proposed SpectraM-PS to demonstrate the efficacy of our dual-scale design with local patch embedding.

For evaluating the contribution of our architecture (SpectraM-PS) and training dataset (PS-Multiplex), we additionally rendered three-channel spectrally multiplexed images representing six distinct surface material categories: (a) uniform, Lambertian; (b) piece-wise uniform, Lambertian; (c) non-uniform, Lambertian; (d) uniform, non-Lambertian; (e) piece-wise uniform, non-Lambertian; and (f) non-uniform, non-Lambertian. In uniform materials, every point on the surface within a scene exhibits the same material properties. For piece-wise uniform materials, each object in a scene is composed of the same material, yet different objects possess distinct materials. Non-uniform materials feature unique PBR textures assigned to each object. The rendering process for these images was identical to that used for the PS-Multiplex datasets in each category. We generated 100 scenes for each surface material category, and MAEs (stds) are averaged over them. The results are presented in Tab.1. In summary, SDM-UniPS [26] trained on our PS-Multiplex dataset demonstrates proper adaptation to the spectrally multiplexed photometric stereo task. Nonetheless, our SpectraM-PS method significantly enhanced reconstruction accuracy, showcasing an architecture-level improvement over SDM-UniPS for the spectrally multiplexed photometric stereo task, where the number of input channels is typically much fewer than that of input images for conventional PS.

Comparative Evaluation on SpectraM14. The results are illustrated in Tabs. 2 to 6 and Fig. 6. Despite the fact that all existing spectrally multiplexed photometric stereo methods assume calibrated light sources and known directional light source conditions, our proposed method significantly outperformed them. This is because most of the real objects used in our experiment are neither Lambertian nor convex, and do not conform to their assumptions. However, our non-physical-based method successfully restored the normals very stably for these objects. Furthermore, our proposed method enabled robust reconstruction for all objects, despite having been trained only with RGB color images. This result supports the efficacy of our approximation. Furthermore, unlike existing methods that suffer from reduced estimation accuracy with increasing spectral crosstalk, our approach demonstrates only minimal performance degradation. Remarkably, our method excels in recovering a more realistic structure with spatially-varying surface materials. This breakthrough implies that our network can effectively achieve dynamic surface reconstruction across video frames in a universal setting. We will detail this groundbreaking application in the next section. Due to space constraints, not all results can be included here. However, all

Table 2: Comparison in condition 1. The values are mean angular errors in degrees.

Method	Object ID														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave.
Ours	**8.6**	**11.2**	**13.0**	**7.3**	**12.2**	**6.1**	**11.5**	**5.0**	**5.7**	**7.5**	**10.3**	**5.1**	**6.1**	**12.2**	**8.9**
Lv *et al.* [46]	20.4	17.0	21.1	13.9	23.1	10.7	21.2	16.6	10.9	15.9	19.0	16.4	13.2	18.6	17.1
Guo *et al.* [17]	22.6	15.2	20.7	13.4	27.1	7.2	31.3	24.8	8.0	18.2	24.5	11.4	10.2	29.3	18.9

Table 3: Comparison in condition 2. The values are mean angular errors in degrees.

Method	Object ID														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave.
Ours	**10.1**	**11.5**	**13.5**	**9.0**	**12.8**	**7.0**	**10.9**	**5.5**	**6.3**	**9.3**	**12.7**	**5.3**	**9.8**	**14.7**	**10.0**
Lv *et al.* [46]	22.8	26.0	27.0	19.4	30.3	19.5	22.1	18.9	14.3	19.8	23.5	20.8	19.0	21.4	21.7
Guo *et al.* [17]	31.1	27.5	29.2	20.1	38.0	19.0	33.4	23.5	13.6	26.6	32.7	14.7	17.1	39.3	25.7

Table 4: Comparison in condition 3. The values are mean angular errors in degrees.

Method	Object ID														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave.
Ours	**12.0**	**12.8**	**15.8**	**11.0**	**16.5**	**5.9**	**12.3**	**9.8**	**6.9**	**9.8**	**14.6**	**6.3**	**7.5**	**20.1**	**11.6**
Lv *et al.* [46]	38.5	34.1	38.2	32.4	40.1	38.2	36.6	29.6	38.2	35.1	38.2	36.6	38.4	30.9	36.0
Guo *et al.* [17]	46.0	42.6	56.3	37.6	57.1	45.6	76.0	48.0	29.9	62.1	49.8	50.2	52.8	73.7	51.7

Table 5: Comparison in condition 4. The values are mean angular errors in degrees.

Method	Object ID														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave.
Ours	**10.9**	**11.3**	**9.7**	**6.3**	**10.8**	**5.3**	**12.4**	**4.3**	**7.8**	**6.6**	**9.1**	**4.0**	**7.4**	**10.6**	**8.4**
Lv *et al.* [46]	24.1	19.8	20.7	12.0	19.8	12.4	15.1	18.7	11.4	17.1	21.2	17.5	19.1	16.7	17.5
Guo *et al.* [17]	30.7	16.0	18.3	9.9	29.6	8.4	23.1	25.7	10.2	14.6	24.5	13.9	29.6	13.8	19.1

Table 6: Comparison in condition 5. The values are mean angular errors in degrees.

Method	Object ID														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ave.
Ours	**10.9**	**11.0**	**9.9**	**7.3**	**11.0**	**4.7**	**13.3**	**4.8**	**7.9**	**6.5**	**9.8**	**4.2**	**7.4**	**10.4**	**8.6**
Lv *et al.* [46]	29.8	26.2	28.9	22.8	28.9	25.8	21.9	24.6	18.7	24.7	27.6	24.8	22.3	27.6	25.0
Guo *et al.* [17]	40.0	27.3	29.2	23.2	33.3	25.0	29.3	30.1	20.0	26.7	31.6	26.3	25.0	27.2	27.8

results are comprehensively presented in the supplementary materials. Additionally, the supplementary materials evaluate the impact of the spatial distribution of light sources on the performance of the proposed method. We also offer an in-depth discussion of each experimental condition therein.

7 Conclusion

In this work, we introduce an innovative approach to spectrally multiplexed photometric stereo under unknown spatial/spectral composition. Turning spectral ambiguity into a benefit, our method allows for the creation of training data without the need for complex multispectral rendering. Our work significantly broadens the scope for dynamic surface analysis, establishing a critical advancement in the utilization of photometric stereo across multiple sectors. Our proposed method exhibits several limitations. Firstly, there is unstable temporal variation in the normal maps reconstructed by our method for dynamic surface reconstruction. This instability arises from factors such as motion blur in certain frames, image noise, or the influence of cast/attached shadows, which become

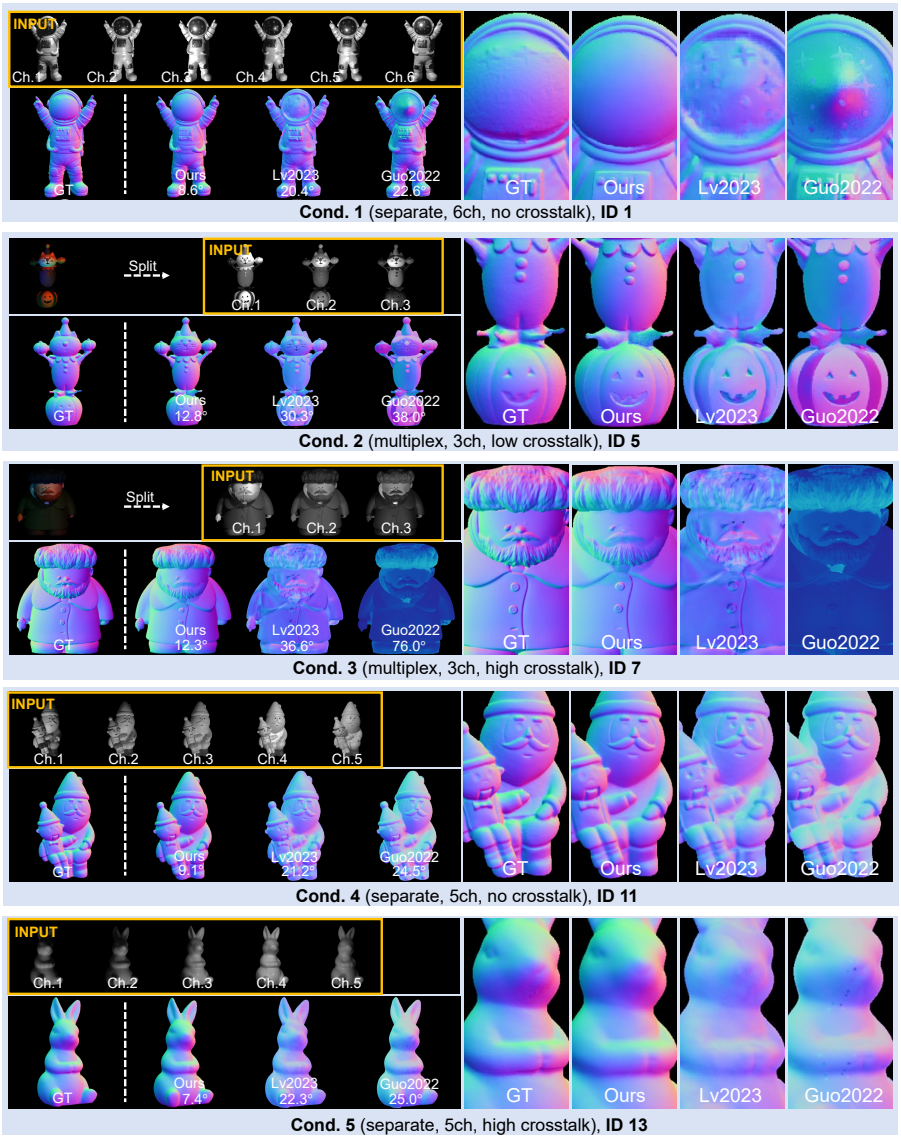

Fig. 6: Evaluation on SpectraM14. Full results are available in the supplementary.

more pronounced compared to conventional photometric stereo methods that utilize numerous images. To recover clean and temporally stable normal maps, we may need to consider temporal consistency and more actively utilize monocular cues. Additionally, while our method targets dynamic surfaces, it currently requires several seconds to up to ten seconds per RGB image, which is far from real-time processing. Considering industrial applications in the future, accelerating the processing speed is a crucial challenge.

References

1. 3D Textures - Free seamless PBR textures with Diffuse, Normal, Displacement, Occlusion, Specularity and Roughness Maps. https://3dtextures.me/, accessed: 2024-03-07 7
2. Adobe Stock. https://stock.adobe.com/ 7, 9
3. AmbientCG - Free Public Domain PBR Materials. https://ambientcg.com/, accessed: 2024-03-07 7
4. Blender. https://www.blender.org/ 9
5. Poliigon - A library of materials, and HDR's for artists including free textures. https://www.poliigon.com/, accessed: 2024-03-07 7
6. Alldrin, N., Mallick, S., Kriegman, D.: Resolving the generalized bas-relief ambiguity by entropy minimization. CVPR (2007) 4
7. Anderson, R., Stenger, B., Cipolla, R.: Color photometric stereo for multicolored surfaces. ICCV (2011) 2
8. Basri, R., Jacobs, D., Kemelmacher, I.: Photometric stereo with general, unknown lighting. International Journal of computer vision **72**(3), 239–257 (2007) 4
9. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**, 151–175 (2010) 7
10. Chakrabarti, A., Sunkavalli, K.: Single-image rgb photometric stereo with spatially-varying albedo. In: 2016 Fourth International Conference on 3D Vision (3DV) (2016) 2
11. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.K.K.: Self-calibrating deep photometric stereo networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8731–8739 (2019) 2, 4, 6
12. Chen, G., Han, K., Wong, K.Y.K.: Ps-fcn: A flexible learning framework for photometric stereo. ECCV (2018) 2, 4, 6
13. Chen, G., Waechter, M., Shi, B., Wong, K.Y.K., Matsushita, Y.: What is learned in deep uncalibrated photometric stereo? In: European Conference on Computer Vision. pp. 745–762. Springer (2020) 4, 6
14. Decker, B., Kautz, J., Mertens, T., Bekaert, P.: Capturing multiple illumination conditions using time and color multiplexing. CVPR (2009) 2
15. Drew, M.S.: Shape from color. Technical Report CSS/LCCR TR 92-07, School of Computing Science, Simon Fraser University, Vancouver, BC (1992) 4
16. Goldman, D., Curless, B., Hertzmann, A., Seitz, S.: Shape and spatially-varying brdfs from photometric stereo. In: ICCV (October 2005) 3, 4
17. Guo, H., Okura, F., Shi, B.: Multispectral photometric stereo for spatially-varying spectral reflectances. IJCV **130**, 2166–2183 (2022) 2, 11, 13
18. Guo, H., Okura, F., Shi, B., Funatomi, T., Mukaigawa, Y., Matsushita, Y.: Multispectral photometric stereo for spatially-varying spectral reflectances: A well posed problem? In: CVPR. pp. 963–971 (2021) 2, 3
19. Hayakawa, H.: Photometric stereo under a light souce with arbitary motion. JOSA **11**(11), 3079–3089 (1994) 4
20. Hernández, C., Vogiatzis, G., Brostow, G., Stenger, B., Cipolla, R.: Non-rigid photometric stereo with colored lights. In: ICCV. pp. 1–8 (2007). https://doi.org/10.1109/ICCV.2007.4408939 2, 4
21. Hertzmann, A., Seitz, S.: Example-based photometric stereo: shape reconstruction with general, varying brdfs. IEEE TPAMI **27**(8), 1254–1264 (2005) 4

22. Hill, B.: Color capture, color management, and the problem of metamerism: does multispectral imaging offer the solution? In: Proc. SPIE. pp. 2–14. SPIE (1999) 2, 8
23. Ikehata, S.: Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In: ECCV (2018) 2, 4, 6
24. Ikehata, S.: Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism. In: BMVC (2021) 4, 6
25. Ikehata, S.: Universal photometric stereo network using global lighting contexts. In: CVPR (2022) 2, 3, 4, 6, 7
26. Ikehata, S.: Scalable, detailed and mask-free universal photometric stereo. In: CVPR (2023) 2, 3, 4, 6, 7, 8, 9, 11, 12
27. Ikehata, S., Aizawa, K.: Photometric stereo using constrained bivariate regression for general isotropic surfaces. In: CVPR (2014) 3
28. Ikehata, S., Wipf, D., Matsushita, Y., Aizawa, K.: Robust photometric stereo using sparse regression. In: CVPR (2012) 3, 4
29. Ikehata, S., Wipf, D., Matsushita, Y., Aizawa, K.: Photometric stereo using sparse bayesian regression for general diffuse surfaces. IEEE TPAMI **36**(9), 1816–1831 (2014) 2, 3
30. Ikehata, S.: Does physical interpretability of observation map improve photometric stereo networks? In: ICIP (2022) 4, 6
31. Ishio, H., Minowa, J., Nosu, K.: Review and status of wavelength-division-multiplexing technology and its application. Journal of lightwave technology **2**(4), 448–463 (1984) 2
32. Ju, Y., Dong, X., Wang, Y., Qi, L., Dong, J.: A dual-cue network for multispectral photometric stereo. Pattern Recognition **100**, 107162 (2020) 2, 4, 6
33. Ju, Y., Qi, L., Zhou, H., Dong, J., Lu, L.: Demultiplexing colored images for multispectral photometric stereo via deep neural networks. IEEE Access **6**, 30804–30818 (2018) 2, 4, 6
34. Ju, Y., Dong, J., Chen, S.: Recovering surface normal and arbitrary images: A dual regression network for photometric stereo. IEEE Transactions on Image Processing **30**, 3676–3690 (2021) 4
35. Kajiya, J.T.: The rendering equation. SIGGRAPH Comput. Graph. **20**(4), 143–150 (1986) 8
36. Kaya, B., Kumar, S., Oliveira, C., Ferrari, V., Van Gool, L.: Uncalibrated neural inverse rendering for photometric stereo of general surfaces. pp. 3804–3814 (2021) 4
37. Kontsevich, L.L., Petrov, A., Vergelskaya, I.: Reconstruction of shape from shading in color images. Journal of the Optical Society of America pp. 1047–1052 (1994) 2, 4
38. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: ICML. pp. 3744–3753 (2019) 7
39. Li, J., Robles-Kelly, A., You, S., Matsushita, Y.: Learning to minify photometric stereo. In: CVPR (2019) 4
40. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: CVPR (2022) 4
41. Li, J., Li, H.: Self-calibrating photometric stereo by neural inverse rendering. In: ECCV (2022) 4
42. Liu, H., Yan, Y., Song, K., Yu, H.: Sps-net: Self-attention photometric stereo network. IEEE Transactions on Instrumentation and Measurement **70**, 1–13 (2021) 4

43. Liu, Z., Mao, H., Chao-Yuan Wu, C.F.: A convnet for the 2020s. In: CVPR (2022) 6
44. Logothetis, F., Budvytis, I., Mecca, R., Cipolla, R.: Px-net: Simple and efficient pixel-wise training of photometric stereo networks. In: CVPR. pp. 12757–12766 (2021) 4
45. Lu, F., Matsushita, Y., Sato, I., Okabe, T., Sato, Y.: Uncalibrated photometric stereo for unknown isotropic reflectances. In: CVPR. pp. 1490–1497 (2013) 4
46. Lv, J., Guo, H., Chen, G., Liang, J., Shi, B.: Non-lambertian multispectral photometric stereo via spectral refectance decomposition. In: IJCAI (2023) 2, 3, 4, 6, 7, 11, 13
47. Mecca, R., Rosman, G., Kimmel, R., Bruckstein, A.: Perspective photometric stereo with shadows. In: Proc. of 4th International Conference on Scale Space and Variational Methods in Computer Vision (2013) 4
48. Mitsubishi Electric Research Laboratories (MERL): MERL BRDF Database. http://www.merl.com/brdf/, accessed: 2024-03-07 7
49. Mo, Z., Shi, B., Lu, F., Yeung, S.K., Matsushita, Y.: Uncalibrated photometric stereo under natural illumination. pp. 2936–2945. IEEE Computer Society (2018) 4
50. Nayatani, Y., Kurioka, Y., Sobagaki, H.: Study on color rendering and metamerism (part 8). Journal of the Illuminating Engineering Institute of Japan **56**(9), 529–536 (1972). https://doi.org/10.2150/jieij1917.56.9_529 2, 8
51. Ozawa, K., Sato, I., Yamaguchi, M.: Single color image photometric stereo for multi-colored surfaces. Computer Vision and Image Understanding **171**, 140–149 (2018) 2
52. Ren, J., Wang, F., Zhang, J., Zheng, Q., Ren, M., Shi, B.: Diligent10^2: A photometric stereo benchmark dataset with controlled shape and material variation. pp. 12581–12590 (June 2022) 3, 9
53. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: International Workshop on Physics Based Vision meets Deep Learning (PBDL) in Conjunction with IEEE International Conference on Computer Vision (ICCV) (2017) 4
54. Shi, B., Matsushita, Y., Wei, Y., Xu, C., Tan, P.: Self-calibrating photometric stereo. In: CVPR (2010) 2, 4
55. Shi, B., Tan, P., Matsushita, Y., Ikeuchi, K.: A biquadratic reflectance model for radiometric image analysis. In: CVPR (2012) 2
56. Shi, B., Wu, Z., Mo, Z., D.Duan, Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In: CVPR (2016) 3, 9
57. Silver, W.M.: Determining shape and reflectance using multiple images. Master's thesis, MIT (1980) 1
58. Taniai, T., Maehara, T.: Neural Inverse Rendering for General Reflectance Photometric Stereo. In: ICML (2018) 4
59. Tiwari, A., Raman, S.: Deepps2: Revisiting photometric stereo using two differently illuminated images. ECCV (2022) 4
60. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017) 7
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 6
62. Vogiatzis, G., Hernandez, C.: Self-calibrated, multi-spectral photometric stereo for 3d face capture. IJCV **56**(97), 91–103 (2012) 2

63. Wang, F., Ren, J., Guo, H., Ren, M., Shi, B.: Diligent-pi: A photometric stereo benchmark dataset with controlled shape and material variation (October 2023) 3, 9
64. Woodham, P.: Photometric method for determining surface orientation from multiple images. Opt. Engg **19**(1), 139–144 (1980) 1, 3
65. Wu, L., Ganesh, A., Shi, B., Matsushita, Y., Wang, Y., Ma, Y.: Robust photometric stereo via low-rank matrix completion and recovery. In: ACCV (2010) 4
66. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV. pp. 418–434 (2018) 6
67. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning. pp. 10524–10533. PMLR (2020) 6
68. Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B.: Gps-net: Graph-based photometric stereo network. NeurIPS (2020) 4
69. Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L.Y., Kot, A.: Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. ICCV (2019) 4