Scalable, Detailed and Mask-Free Universal Photometric Stereo

Satoshi Ikehata
National Institute of Informatics (NII)
sikehata@nii.ac.jp

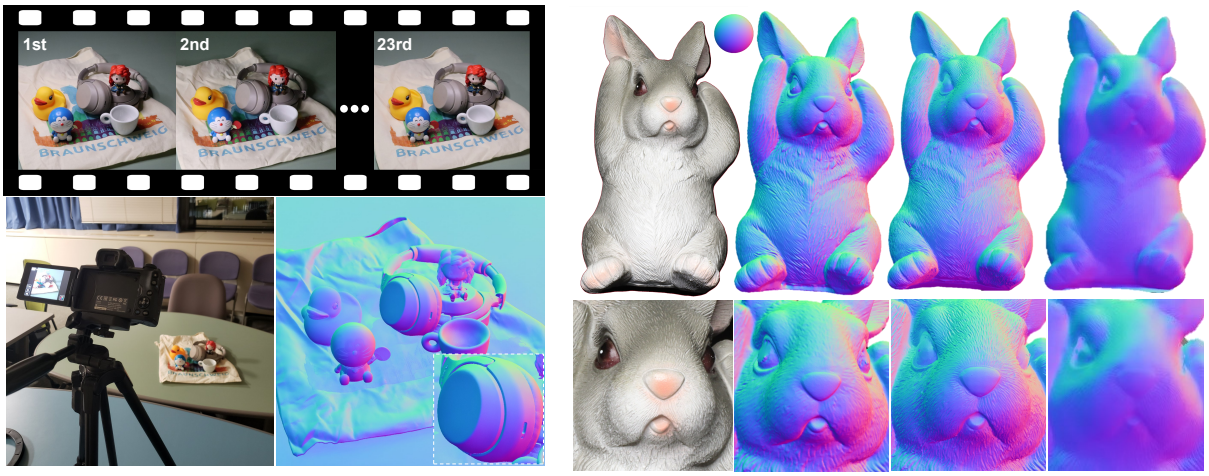


Figure 1. Given multiple images under unknown spatially-varying illuminations, our method can recover the detailed surface normal map of non-convex, non-Lambertian surfaces (Left). Our method even surpasses the level of detail provided by consumer 3-D scanners (Right).

Abstract

In this paper, we introduce SDM-UniPS, a groundbreaking Scalable, Detailed, Mask-free, and Universal Photometric Stereo network. Our approach can recover astonishingly intricate surface normal maps, rivaling the quality of 3D scanners, even when images are captured under unknown, spatially-varying lighting conditions in uncontrolled environments. We have extended previous universal photometric stereo networks to extract spatial-light features, utilizing all available information in high-resolution input images and accounting for non-local interactions among surface points. Moreover, we present a new synthetic training dataset that encompasses a diverse range of shapes, materials, and illumination scenarios found in real-world scenes. Through extensive evaluation, we demonstrate that our method not only surpasses calibrated, lighting-specific techniques on public benchmarks, but also excels with a significantly smaller number of input images even without object masks.

1. Introduction

Photometric stereo [52] aims to deduce the surface normal map of a scene by analyzing images captured from a fixed perspective under diverse lighting conditions. Until very recently, all photometric stereo methods assumed their specific lighting conditions, which led to limitations in their applicability. For instance, methods that assumed directional lighting conditions (*e.g.*, [20, 24, 25]) were unsuitable under natural illumination, and vice versa (*e.g.*, [15, 38]).

To overcome this limitation, the "universal" photometric stereo method (UniPS) [22] has been introduced, designed to operate under unknown and arbitrary lighting conditions. In contrast to prior uncalibrated photometric stereo methods [7, 9, 27], which assumed specific physically-based lighting models, this method encodes a non-physical feature at each pixel for representing spatially-varying illumination, which is served as a substitute for physical lighting parameters within the calibrated photometric stereo network [21]. This method has taken the first step towards dealing with unknown, spatially-varying illumination that none of the existing methods could handle. However, the surface normal map recovered by UniPS, while not entirely inaccurate, appears blurry and lacks fine detail (see the top-right corner of Fig. 1). Upon investigation, we pinpointed three fundamental factors contributing to the subpar reconstruction performance. Firstly, extracting illumination

Supported by JSPS KAKENHI Grant Number 22K17919.

features (*i.e.*, global lighting contexts) from downsampled images caused a loss of information at higher input resolutions and produced blurry artifacts. Secondly, UniPS employs a pixel-wise calibrated photometric stereo network to predict surface normals using illumination features, which leads to imprecise overall shape recovery. Although pixel-wise methods [20,21,25] offer advantages in capturing finer details compared to image-wise methods [8, 29, 49], they suffer from an inability to incorporate global information.

Lastly, the third issue lies in the limited variety of shape, material, and illumination conditions present in the training data, which hampers its capacity to adapt to a diverse range of real-world situations. This limitation primarily stems from the fact that current datasets (*i.e.*, PS-Wild [22]) do not include renderings under light sources with high-frequency components focused on specific incident angles, such as point or directional sources. Consequently, the method exhibits considerable performance degradation when exposed to directional lighting setups like DiLiGenT [46], as will be demonstrated later in this paper.

In this paper, we present a groundbreaking photometric stereo network, the Scalable, Detailed, and Mask-Free Universal Photometric Stereo Network (SDM-UniPS), which recovers normal maps with remarkable accuracy from images captured under extremely uncontrolled lighting conditions. As shown in Fig. 1, SDM-UniPS is *scalable*, enabling the generation of normal maps from images with substantially higher resolution (*e.g.*, 2048x2048) than the training data (*e.g.*, 512x512); it is *detailed*, providing more accurate normal maps on DiLiGenT [46] with a limited number of input images than most existing orthographic photometric stereo techniques, including calibrated methods, and in some cases, surpassing 3D scanners in detail; and it is *mask-free*, allowing for application even when masks are absent, unlike many conventional methods. Our technical novelties include:

1. The development of a *scale-invariant spatial-light feature encoder* that efficiently extracts illumination features while utilizing all input data and maintaining scalability with respect to input image size. Our encoder, based on the "split-and-merge" strategy, accommodates varying input image sizes during training and testing without sacrificing performance.

2. The development of a surface normal decoder utilizing our novel *pixel-sampling transformer*. By randomly sampling pixels of fixed size, we simultaneously predict surface normals through non-local interactions among sampled pixels using Transformers [51], effectively accounting for global information.

3. The creation of a new synthetic training dataset, comprising multiple objects with diverse textures within a scene, rendered under significantly varied lighting conditions that include both low and high-frequency illuminations.

We believe that the most significant contribution is the extraordinary time savings from data acquisition to normal map recovery compared to existing photometric stereo algorithms requiring meticulous lighting control, even in the uncalibrated setup. This progress allows photometric stereo to be executed at home, literally "in the wild" setup.

2. Related Works

In this section, we provide a succinct overview of photometric stereo literature focusing on the single orthographic camera assumption. Alternative setups (*e.g.*, perspective, multi-view cameras) are beyond the scope of this work.

Optimization-based Approach: The majority of photometric stereo methods assume calibrated, directional lighting following Woodham [52] and optimize parameters by inversely solving a physics-based image formation model. This approach can be further categorized into robust methods, where non-Lambertian components are treated as outliers [24, 39, 53, 59]; model-based methods, which explicitly account for non-Lambertian reflectance [13,23,45]; and example-based methods [17, 19, 47] that leverage the observations of known objects captured under identical conditions as the target scene. The uncalibrated task is akin to the calibrated one, but with unknown lighting parameters. Until recently, most uncalibrated photometric stereo algorithms assumed Lambertian integrable surfaces and aimed to resolve the General Bas-Relief ambiguity [4, 11, 12, 16, 36,42,44,55]. In contrast to these works, photometric stereo under natural lights has also been explored, wherein natural illumination is approximated using spherical harmonics [5,15], dominant sun lighting [3,18], or equivalent directional lighting [14, 38]. Although most optimization-based methods do not require external training data, they are fundamentally limited in handling global illumination phenomena (*e.g.*, inter-reflections) that cannot be described by the predefined point-wise image formation model.

Learning-based Approach: Learning-based methods are effective in addressing complex phenomena that are challenging to represent within simple image formation models. However, the first photometric stereo network [43] necessitated consistent lighting conditions during both training and testing. To address this limitation, various strategies have been investigated, such as observation maps [20, 35], set-pooling [8, 26], graph-convolution [58], and self-attention [21, 31]. Furthermore, researchers have explored uncalibrated deep photometric stereo networks [7,9,27,50], where lighting parameters and surface normals are recovered sequentially. Self-supervised neural inverse rendering methods have been developed without the need for exter-

nal data supervision. Taniai and Maehara [49] used neural networks instead of parametric physical models, with images and lighting as input. This work was expanded by Li and Li [29, 30], who incorporated recent neural coordinate-based representations [37]. However, despite their tremendous efforts, these methods are designed to work with only single directional light source and have limited ability to generalize to more complex lighting environments.

Universal Photometric Stereo Network: The universal photometric stereo network (UniPS) [22] was the first to eliminate the prior lighting model assumption by leveraging a non-physical lighting representation called global lighting contexts. These global lighting contexts are recovered for each lighting condition through pixel-wise communication of hierarchical feature maps along the light-axis using Transformers [51]. During surface normal prediction, a single location is individually selected, and the network aggregates all the global lighting contexts (bilinearly interpolated from the canonical resolution) and raw observations at the location under different lighting conditions to pixel-wise predict the surface normal. This method introduced two strategies to handle high-resolution images: down-sampling images to the canonical resolution for recovering global lighting contexts, and employing pixel-wise surface normal prediction. Although these two concepts contributed to the scalability of image size, they resulted in performance degradation due to the loss of input information and the absence of a non-local perspective, as previously discussed.

Our work draws inspiration from [22] and shares some fundamental ideas, particularly the use of Transformers [51] for communicating and aggregating features along the light-axis. However, our method diverges from [22] by fully utilizing input information in a non-local manner, which leads to a significant enhancement in reconstruction quality.

3. Method

We target the challenging universal photometric stereo task, which was recently introduced in [22]. Unlike prior calibrated and uncalibrated tasks, the universal task makes no assumptions about surface geometry, material properties, or, most importantly, lighting conditions. The objective of this task is to recover a normal map $N \in \mathbb{R}^{H\times W\times 3}$ from images $I_k \in \mathbb{R}^{H\times W\times 3}; k \in 1,\dots,K$ captured under K unknown lighting conditions using an orthographic camera. Optionally, an object mask $M \in \mathbb{R}^{H\times W}$ may be provided.

Our method (SDM-UniPS) is illustrated in Fig. 2. Given pre-processed images and an optional object mask, feature maps for each lighting condition are extracted through interactions along the spatial and light axes (*i.e.*, the scale-invariant spatial-light feature encoder). We then randomly sample locations from the coordinate system of the input image and bilinearly interpolate features at these locations. Features and raw observations at each location are aggregated pixel-wise, and surface normals are recovered from the aggregated features after non-local spatial interaction among them (*i.e.*, the pixel-sampling Transformer). In line with [22], we focus on describing high-level concepts rather than providing detailed explanations for the sake of clarity. Refer to the appendix for a comprehensive description of the network architectures.

3.1. SDM-UniPS

Pre-processing: As in [22], we resize or crop input images to a resolution (R) that is divisible by 32, which is accepted by most hierarchical vision backbones. To ensure that image values are within a similar range, each image is normalized by a random value between its maximum and mean.

Scale-invariant Spatial-light Feature Encoder: After the pre-processing, we extract feature maps from images and an optional object mask through the interaction along both spatial and light axes. Following the basic framework in [22], each image and object mask [1] are concatenated to form a tensor $O_k \in \mathbb{R}^{R\times R\times 4}$, which is then input to the common vision backbone [32–34] to extract hierarchical feature maps $B_k^s \in \mathbb{R}^{\frac{R}{S_s}\times\frac{R}{S_s}\times C_s}, ; s \in 1,2,3,4$. Here, $S_s \in 4,8,16,32$ represents the scale of the s-th feature map, and C_s is the dimension of features at that scale. For each feature scale, features from different tensors at the same pixel interact with each other along the light-axis using naïve Transformers [51]. Finally, hierarchical feature maps are fused to $\mathcal{F}_k \in \mathbb{R}^{\frac{R}{4}\times\frac{R}{4}\times C_\mathcal{F}}$ using the feature pyramid network [56], where $C_\mathcal{F}$ is the output feature dimension. Note that, unlike [22], we used a varying number of Transformer blocks at each hierarchy scale (*i.e.*, the number of blocks changes from [1,1,1,1] to [0,1,2,4]) so that the deeper features interact more than the shallow ones.

In UniPS [22], images and a mask are down-sampled to a *canonical resolution* before being input to the backbone network. This resolution must be constant and sufficiently small (*e.g.*, 256x256) to prevent excessive memory consumption during feature extraction, particularly when dealing with high-resolution input images. Additionally, using a constant resolution ensures that tensors of the same shape are fed to the backbone, which helps to avoid significant performance degradation due to a large discrepancy in input tensor shapes between training and testing. Consequently, down-sampling leads to the loss of much information in the input images, resulting in a blurry normal map recovery.

To address it, we propose a *scale-invariant spatial-light feature encoder* designed to maintain a consistent, small input resolution for the backbone network while preserving information from input images. Specifically, instead of downsampling, we suggest *splitting* the input tensor into

[1] Without a mask, a matrix with all values set to one is concatenated.

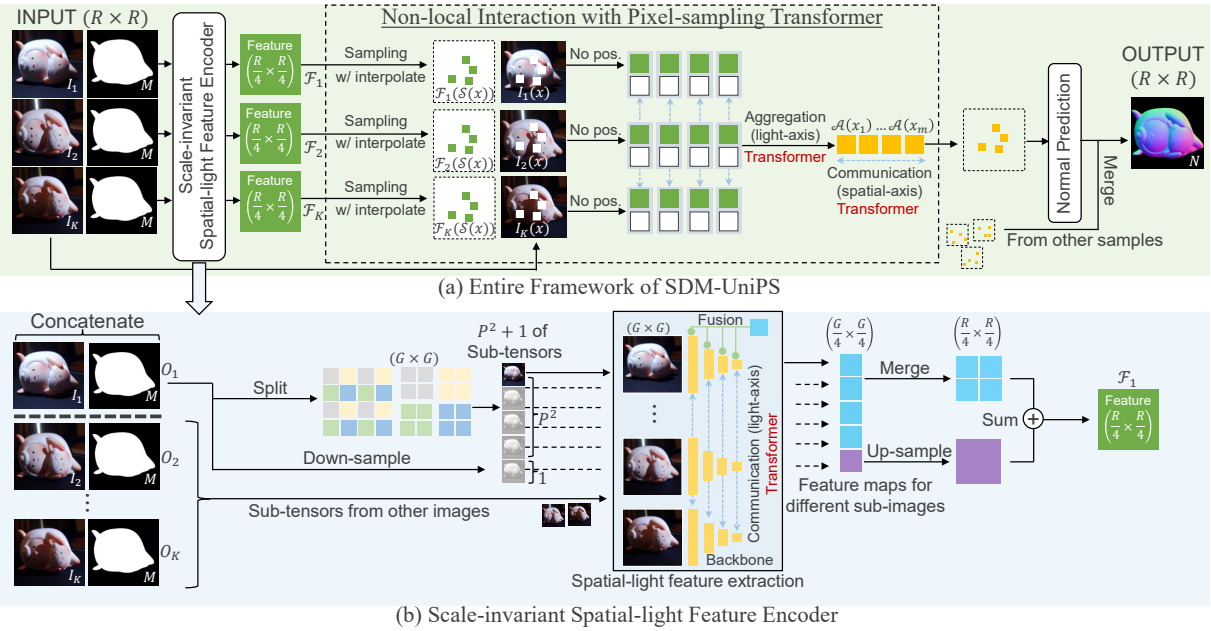


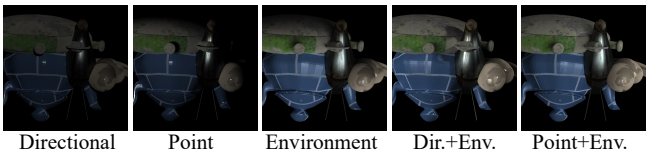
Figure 2. Our entire framework is illustrated in (a). Given multiple images and an object mask (optional), the scale-invariant spatial-light encoder (detailed in (b)) extracts a feature map for each image. The surface normal vectors are independently recovered at each of pixel samples (*i.e.*, 2048) after the non-local interaction among aggregated features from interpolated feature maps and raw observations.

non-overlapping sub-tensors with a constant, small resolution. In greater detail, we decompose O into P^2 sub-tensors of size $G \times G$ ($G = 256$ in our implementation, $P \triangleq R/G$) by taking a single sample from every $P \times P$ pixel and stacking them as sub-tensors, as illustrated in Fig. 2. Each sub-tensor encompasses the entire portion of the original tensor but is slightly shifted. All sub-tensors are processed independently through the same spatial-light feature encoder and subsequently merged back into a tensor of size $(\frac{R}{4} \times \frac{R}{4} \times C_{\mathcal{F}})$. The combined feature maps from the sub-tensors retain all input information since no downsampling occurred. However, the absence of interaction among different sub-tensors leads to significant block artifacts, particularly when P is large. To mitigate this, another feature map encoded from the naively downsized image is added[2] to the merged feature maps, promoting interaction among sub-tensors. Optionally, when P is larger than 4, we apply depth-wise Gaussian filtering (*i.e.*, kernel size is P-1) to the feature maps to further enhance the interaction. Finally, we obtain the scale-invariant spatial-light feature maps $\mathcal{F}_k \in \mathbb{R}^{\frac{R}{4} \times \frac{R}{4} \times C_{\mathcal{F}}}$ for every lighting condition.

Non-local Interaction with Pixel-sampling Transformer: Given the scale-invariant spatial-light feature maps $\mathcal{F}_k$ and input images I_k, the surface normal is recovered after pixel-wise feature aggregation along the light-axis (*i.e.*, the light channel shrinks from K to 1). Feature aggregation under different lighting conditions is a fundamental step in photometric stereo networks, and various strategies have been studied, such as observation maps [20, 35], max-pooling [8, 9], graph-convolution [58], and self-attention [21, 22]. We utilize the Transformer model with self-attention [51] as in the encoder following UniPS [22]. UniPS directly predicted surface normals from pixel-wise aggregated feature vectors, following other pixel-wise methods [20, 21, 25], without considering non-local interactions. However, aggregated features lose lighting-specific information, naturally obtaining lighting-invariant representations more related to surface attributes than those before aggregation. In traditional physics-based vision tasks, common constraints including isotropy [4], reciprocity symmetry [48], reflectance monotonicity [6], sparse reflectance basis [13], and surface integrability [41] are mostly shared on the surface, not limited to a single surface point. Thus, considering non-local interactions of aggregated features at multiple surface points is crucial in physics-based tasks.

Applying image-wise neural networks like CNNs on the aggregated feature map demands enormous computational cost for large output resolutions (*e.g.*, 2048×2048), and risks compromising output normal map details. To address these issues, we draw inspiration from recent Transformers on 3-D points [54, 61] and apply a Transformer on a fixed number (m) of *pixel samples* (*e.g.*, $m = 2048$) from random locations in the input coordinate system. We term this the *pixel-sampling Transformer*. Unlike image-based approaches, pixel-sampling Transformer's memory consumption is constant per sample set, scaling to arbitrary image

[2] Concatenation is also possible, but it did not improve the results despite increased memory consumption.



Directional Point Environment Dir.+Env. Point+Env.

Figure 3. Examples in PS-Mix under different lighting conditions.

sizes. Moreover, by applying the Transformer to a randomly sampled set of locations, local interactions that may lead to over-smoothing of feature maps (*e.g.*, in CNNs) are almost entirely eliminated.

Concretely, given m random pixels $x_{i=1,\dots,m}$ from the masked region of the input coordinate system, we interpolate features at those pixels as $\mathcal{F}_{1,\dots,K}(\mathcal{S}(x_i))$, where $\mathcal{S}$ is the bilinear interpolation operator. Then, interpolated features are concatenated with corresponding raw observations $I_{1,\dots,K}(x_i)$ and aggregated to $\mathcal{A}(x_i)$ with pooling by multi-head attention (PMA) [28], as in [22]. Given aggregated features at different pixels in the same sample set, we apply another naïve Transformer [51] to perform non-local interactions. Since the goal of this process is to consider surface-level interactions based on physical attributes, pixel coordinate information is unnecessary. Thus, we *don't* apply position embeddings to samples, unlike most existing visual Transformer models (e.g. [10,33]), allowing the samples to propagate their aggregated features without location information.

After the non-local interaction, we apply a two-layer MLP to predict surface normals at sampled locations. Finally, surface normals for each set are merged to obtain the final surface normal map at the input image resolution. This pixel-sampling Transformer approach facilitates non-local interactions while maintaining computational efficiency and preserving output normal map details, making it suitable for physics-based tasks with high-resolution images.

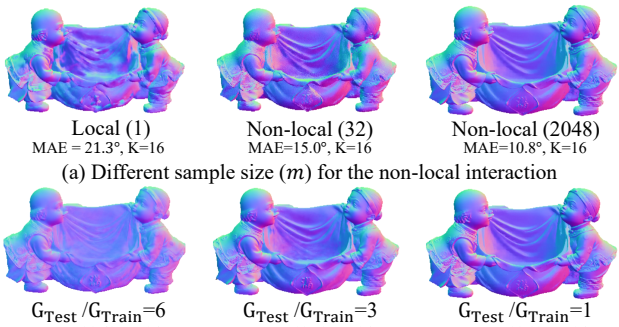
3.2. PS-Mix Dataset

To train their universal photometric stereo network, Ikehata [22] presented the PS-Wild training dataset, which rendered more than 10,000 scenes with commercial AdobeStock 3-D assets [1]. One of the issues in PS-Wild is that each scene consists of only a single object of uniform material. Furthermore, the environment lighting used for rendering scenes in [22] rarely has high-frequency illumination (*e.g.*, a single point light source); therefore, the rendered images are biased towards low-frequency lighting conditions.

In this paper, we create a new training dataset that solves the issues in the PS-Wild training dataset. Instead of putting a single object of uniform material in each scene, we put multiple objects that overlap with each other in the same scene and give them different materials. To ensure that the material category is diverse in a scene, we manually categorized 897 texture maps in the AdobeStock material assets into 421 diffuse, 219 specular, and 257 metallic textures.

Table 1. Ablation study on PS-Wild-Test [22].

Method	Training	Dir.	HDRI	Dir.+HDRI
I22 (UniPS) [22]	PS-Wild	17.0	14.5	13.8
Only Local (baseline)	PS-Mix	8.4	14.7	11.8
+Non-local (32)	PS-Mix	7.8	14.9	10.8
+Non-local (128)	PS-Mix	6.2	13.0	8.9
+Non-local (512)	PS-Mix	5.8	12.4	8.2
+Non-local (2048)	PS-Mix	5.7	12.2	8.0
+Non-local (20480)	PS-Mix	5.7	12.3	8.0
+Scale-invariant Enc.	PS-Mix	4.8	11.1	7.5



(a) Different sample size (m) for the non-local interaction

(b) Different input resolutions between training and test

Figure 4. (a) Comparison of different sample size (m) for the non-local interaction in the normal prediction. (b) Comparison of different input resolutions to the encoder between training and test.

For each scene, we randomly select four objects from 410 AdobeStock 3-D models and assign three textures from all three material categories and randomly choose one for each object. Furthermore, to make the lighting conditions more diverse, instead of using only environment lighting to render images, we use five types of light source configurations and mix them to render one scene; (a) environment lighting, (b) single directional lighting, (c) single point lighting, (d) (a)+(b), and (e) (a)+(c). The direction and position of light sources are randomly assigned within the valid range of parameters[3]. We followed PS-Wild [22] for other rendering techniques (*e.g.*, auto-exposure, object scale adjustment). Our dataset consists of 34,921 scenes, and each scene is rendered to output 10 of 16-bit, 512×512 images. In Fig. 3, we show sample images under each lighting condition for the same scene.

4. Results

Training Details: Our network was trained on the PS-Mix dataset from scratch using the AdamW optimizer and a step decay learning rate schedule (×0.8 every ten epochs) with learning-rate warmup during the first epoch. A batch size of 8, an initial learning rate of 0.0001, and a weight

[3]Light directions are selected from the upper unit hemisphere, and point light positions are selected inside the hemisphere.

Table 2. Evaluation on DiLiGenT [46] (Mean Angular Errors in degrees). All 96 images were used except where K is shown.

Method	Approach	Task	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Ave.
Woodham [52]	Self-Sup.	Calibrated	4.1	8.4	14.9	8.4	25.6	18.5	30.6	8.4	14.7	19.8	15.3
IW14 [25]	Self-Sup.	Calibrated	2.0	4.8	8.4	5.4	13.3	8.7	18.9	6.9	10.2	12.0	9.1
IA14 [23]	Self-Sup.	Calibrated	3.3	7.1	10.5	6.7	13.1	9.7	26.0	6.6	8.8	14.2	10.6
I18 [20]	Supervised	Calibrated	2.2	4.1	7.9	4.6	8.0	7.3	14.0	5.4	6.0	12.6	7.2
CW20 [9]	Supervised	Calibrated	2.7	7.7	**7.5**	4.8	6.7	7.8	12.4	6.2	7.2	10.9	7.4
LB21 [35]	Supervised	Calibrated	2.0	**3.5**	7.6	**4.3**	4.7	**6.7**	13.3	4.9	5.0	9.8	6.2
LL22a [29]	Sup.+Self-Sup.	Calibrated	2.4	3.6	8.0	4.9	4.7	**6.7**	14.9	6.0	5.0	8.8	6.5
CH19 [7]	Supervised	Uncalibrated	2.8	6.9	9.0	8.1	8.5	11.9	17.4	8.1	7.5	14.9	9.5
CW20 [9]	Supervised	Uncalibrated	2.5	5.6	8.6	7.9	7.8	9.6	16.2	7.2	7.1	14.9	8.7
KK21 [27]	Sup.+Self-Sup.	Uncalibrated	3.8	6.0	13.1	7.9	10.9	11.9	25.5	8.8	10.2	18.2	11.6
LL22b [30]	Sup.+Self-Sup.	Uncalibrated	**1.2**	3.8	9.3	4.7	5.5	7.1	14.6	6.7	6.5	10.5	7.0
TR22 [50] (K=2)	Self-Sup.	Uncalibrated	6.3	9.7	14.5	9.9	11.1	14.2	26.1	10.7	12.1	19.9	13.4
I22 (UniPS) [22]	Supervised	Universal	4.9	9.1	19.4	13.0	11.6	24.2	25.2	10.8	9.9	18.8	14.7
Ours	Supervised	Universal	1.5	3.6	**7.5**	5.4	**4.5**	8.5	**10.2**	**4.7**	**4.1**	**8.2**	**5.8**
Ours (K=64)	Supervised	Universal	1.5	3.6	7.6	5.5	4.6	8.6	10.2	4.7	4.1	8.3	5.9
Ours (K=32)	Supervised	Universal	1.5	3.6	7.7	5.5	4.7	8.6	10.4	4.8	4.2	8.4	5.9
Ours (K=16)	Supervised	Universal	1.5	3.8	7.7	6.0	4.8	8.5	10.8	4.9	4.4	8.7	6.1
Ours (K=8)	Supervised	Universal	1.6	4.0	8.2	6.3	5.2	8.4	11.5	5.2	4.8	9.4	6.5
Ours (K=4)	Supervised	Universal	1.7	4.1	10.0	8.6	6.3	9.0	14.1	6.1	5.9	11.4	7.7
Ours (K=2)	Supervised	Universal	1.9	6.8	14.4	13.6	8.3	12.8	21.2	9.0	9.2	16.9	11.4

decay of 0.05 were used. The number of input training images was randomly selected from 3 to 6 for each batch [4]. In our work, we chose ConvNeXt-T [34] as our backbone due to its simplicity and efficiency, which is better than recent ViT-based architectures [10, 32, 33] with comparable performance. The training loss was the MSE loss, which computes the ℓ_2 errors between the prediction and ground truth of surface normal vectors. Additional information, such as network architectures and feature dimensions, is provided in the appendix.

Evaluation and Time: The accuracy is evaluated based on the mean angular errors (MAE) between the predicted and true surface normal maps, measured in degrees. Training is conducted on four NVIDIA A100 cards for roughly three days. The inference time of our method depends on the number and resolution of input images. In the case of 16 input images at a resolution of 512×512, it takes a few seconds excluding I/O on a GPU. While the computational cost will vary almost linearly with the number of images, this is significantly more efficient than recent neural inverse rendering-based methods [29, 30, 40, 60].

4.1. Ablation Study

Firstly, we perform ablation studies to evaluate the individual contributions of our scale-invariant spatial-light feature encoder and non-local interaction with pixel-sampling transformer across varying sample sizes. To quantitatively compare performance under various lighting conditions, we utilize the PS-Wild-Test dataset [22], which contains 50 synthetic scenes rendered under three distinct lighting setups: directional, environmental, and a mixture of both. In Table 1 and Fig. 4, we compare our method with different configurations against [22]. Note that without the scale-invariant encoder and non-local interaction (*i.e.*, the baseline), our method is nearly equivalent to [22], except for some minor differences (*e.g.*, backbone architecture, number of Transformer blocks in the encoder). We observe that the baseline method trained on our PX-Mix dataset improves performance for scenes under directional lighting, suggesting that one of the primary reasons why [22] was ineffective under directional lights was due to bias in the PS-Wild dataset. Accounting for non-local interaction of aggregated features enhances reconstruction accuracy, even with a small number of samples (*e.g.*, m=32), as clearly illustrated in Fig. 4-(top). Although accuracy improved as the number of samples increased, as expected, performance gains plateaued beyond a certain number (*i.e.*, m=2048). The efficacy of the scale-invariant spatial-light feature encoder was also confirmed. In Fig. 4-(bottom), we observed that a significant difference in input resolution to the backbone between training and testing led to substantial performance degradation, which further validates the advantage of our method that maintains a constant input tensor shape.

[4] Six is the maximum number that can fit on our GPU.

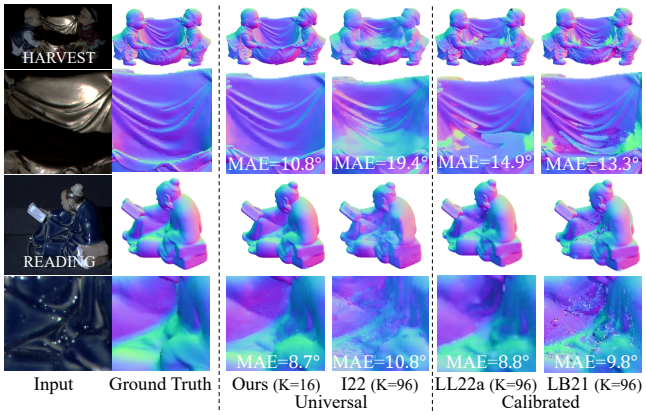


Figure 5. Results for objects under a single directional lighting condition, including object masks.

4.2. Evaluation under Directional Lighting

DiLiGenT Evaluation: We first evaluate our method on the DiLiGenT benchmark [46]. Each dataset provides 96 612x512 16-bit HDR images, captured under known single directional lighting. The object mask and true surface normal map are available. In addition to UniPS [22], we also compare our method with calibrated [9,20,23,25,29,35,52] and uncalibrated [7, 9, 27, 30, 50] photometric stereo algorithms specifically designed for single directional lighting. Calibrated methods include both pixelwise [20, 23, 25, 35, 52] and image-wise [9, 29] approaches. All uncalibrated methods are image-wise. We consider [27,29,30] as a combination of supervised and unsupervised learning, as pretrained models were used as a starting point for lighting prediction. To evaluate the valid number of input images, we compare our method with different numbers of input images (results are averaged over 10 random trials).

The results are illustrated in Table 2. Impressively, our method, which does not assume a specific lighting model, outperforms state-of-the-art calibrated methods designed for directional lights (LB21 [35], LL22a [29]). Furthermore, unlike conventional photometric stereo methods, the proposed method does not experience significant performance degradation even when the number of input images is reduced; it maintains state-of-the-art results even with only 8 images. The proposed method ($K = 2$) also surpasses TR22 [50], which is specialized for two input images.

Recovered normal maps of HARVEST and READING are shown in Fig. 5. These objects are considered the most challenging in the benchmark due to their highly non-convex geometry. As expected, the state-of-the-art pixel-wise calibrated method (LB21 [35]) can recover finer surface details, while the state-of-the-art image-wise calibrated method (LL22a [29]) can recover more globally consistent results. However, both of them struggle to recover the non-convex parts of the objects accurately. On the other hand, our method can recover both surface details and overall shape without apparent difficulty, even with a much smaller number of images (*i.e.*, K=16). As expected, the performance of I22 [22] is severely lacking.

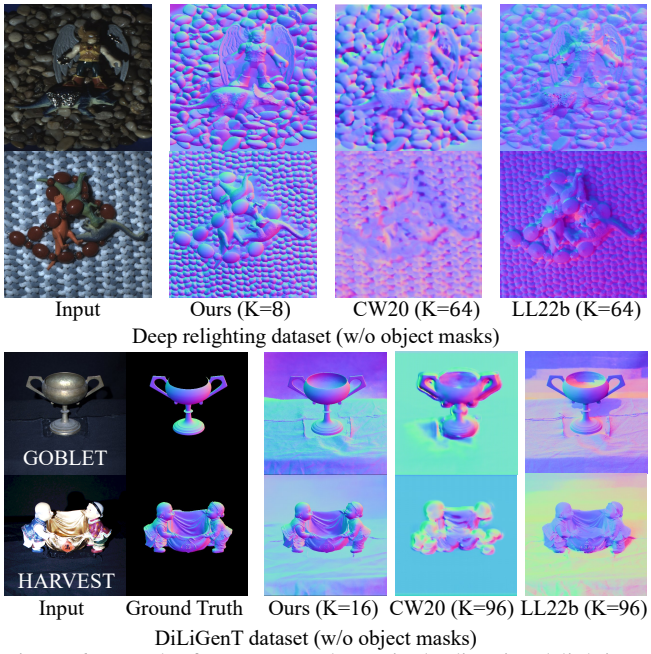


Figure 6. Results for scenes under a single directional lighting condition, excluding object masks.

Evaluation without Object Mask: To demonstrate that our method does not require an object mask, we applied it to two real scenes from a deep relighting work [57], each containing 530 8-bit integer images at a resolution of 512x512, captured under unknown single directional lighting using a gantry-based acquisition system. The object mask and true surface normal map are unavailable. We compared our method with state-of-the-art uncalibrated methods (CW20 [9] and LL22b [30]) and displayed the results in Fig. 6 (top). Unlike the uncalibrated methods that struggled to recover accurate lighting directions, our proposed method successfully captured object boundaries without masks, even in complex scenes with significant global illumination effects, and consistently recovered normals across the entire image. We further evaluated our method on DiLiGenT scenes without masks, as illustrated in Fig. 6 (bottom). While existing methods that assume an object mask produced highly inaccurate surface normal maps, our proposed method recovered more plausible normals with fewer images (*i.e.*, K=16 vs K=96).

4.3. Evaluation under Spatially-varying Lighting

Our method is evaluated on challenging scenes with spatially-varying lighting conditions, comparing it to the first universal network (UniPS) [22] and a state-of-the-art uncalibrated photometric stereo method (GM21) [14] on

Figure 7. Qualitative comparison on images under spatially-varying lighting conditions with object masks [22].

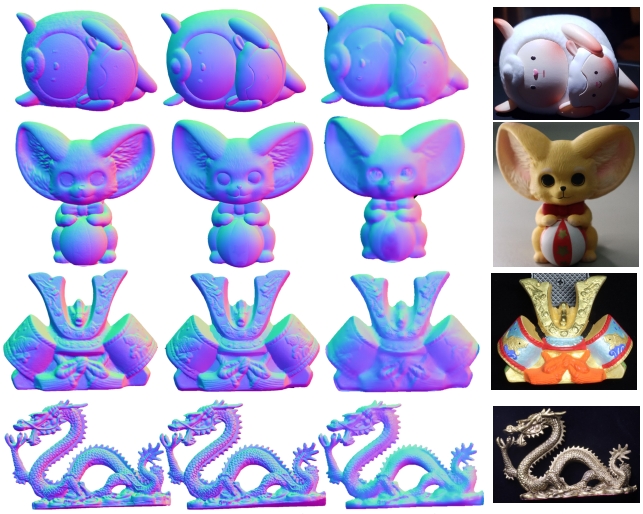


Figure 8. Qualitative comparison with 3-D scans.

Figure 9. Surface normal recovery from images under spatially-varying lighting conditions without object masks.

a dataset provided by [22]. We test three objects (Apple, Grape, and Owl). While GM21 [14] fails and I22 [22] loses details, our method, using a scale-invariant spatial-light feature encoder and non-local interaction, produces accurate results.

In Fig. 8, we subjectively compare our method using four objects with normal maps obtained from a 3D scanner. We align the scanned normal map to the image using MeshLab's mutual information registration filter [2], as in [46]. Our method recovers higher-definition surface normal maps than the 3D scanner (EinScan-SE) and performs well regardless of surface material. Photometric stereo performance improves with increased digital camera resolution, suggesting that 3D scanners may struggle to keep up.

Lastly, we demonstrate surface normal prediction for complex non-convex scenes without masks under challenging lighting conditions in Figure 9. We apply our method to three extremely challenging datasets: School Desk, Coins and Keyboard, and Sweets. School Desk is a complex scene with simple objects, non-uniform lighting, and cast shadows, making surface normal map recovery difficult. Coins and Keyboard features multiple planar objects of various materials. Sweets is a challenging scene with abundant inter-reflections and cast shadows. As demonstrated, the proposed method successfully recovers uniform surface normals, largely unaffected by shadows, and effectively reconstructs the surface micro-shape, demonstrating its scalability and detail preservation.

5. Conclusion

In this paper, we presented a scalable, detailed, and mask-free universal photometric stereo method. We demonstrated that the proposed method outperforms most calibrated and uncalibrated methods in the DiLiGenT benchmark. In addition, the comparison with the only existing method [22] for the universal task showed a significant improvement over it.

However, several challenges still remain. Firstly, although we have observed that the proposed method works robustly for versatile lighting conditions, we found that our method is not very effective when the lighting variations are minimal. Secondly, the proposed method can easily be extended beyond normal map recovery by replacing the loss and data. In reality, we have attempted to output BRDF parameters for materials. However, due to fundamental ambiguities, it is difficult to evaluate the recovered BRDF parameters. Please see the appendix for further discussions of these limitations and a variety of additional results to better understand this study.

References

[1] Adobe Stock. https://stock.adobe.com/. 5

[2] Meshlab. https://www.meshlab.net/. 8

[3] J. Ackermann, F. Langguth, S. Fuhrmann, and M. Goesele. Photometric stereo for outdoor webcams. In *CVPR*, 2012. 2

[4] N. Alldrin, S. Mallick, and D. Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. *CVPR*, 2007. 2, 4

[5] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of computer vision*, 72(3):239–257, 2007. 2

[6] M. Chandraker and R. Ramamoorthi. What an image reveals about material reflectance. In *ICCV*, 2011. 4

[7] G. Chen, K. Han, B. Shi, Y. Matsushita, and K. K. K. Wong. Self-calibrating deep photometric stereo networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8731–8739, 2019. 1, 2, 6, 7

[8] G. Chen, K. Han, and K-Y. K. Wong. Ps-fcn: A flexible learning framework for photometric stereo. *ECCV*, 2018. 2, 4

[9] 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. 1, 2, 4, 6, 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 6

[11] Ondřej Drbohlav and Radim Šára. Specularities reduce ambiguity of uncalibrated photometric stereo. In *ECCV*, pages 46–60, 2002. 2

[12] P. Favaro and T. Papadhimitri. A closed-form solution to uncalibrated photometric stereo via diffuse maxima. In *CVPR*, 2012. 2

[13] D. Goldman, B. Curless, A. Hertzmann, and S. Seitz. Shape and spatially-varying brdfs from photometric stereo. In *ICCV*, October 2005. 2, 4

[14] Heng Guo, Zhipeng Mo, Boxin Shi, Feng Lu, Sai Kit Yeung, Ping Tan, and Yasuyuki Matsushita. Patch-based uncalibrated photometric stereo under natural illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2, 7, 8

[15] Bjoern Haefner, Zhenzhang Ye, Maolin Gao, Tao Wu, Yvain Quéau, and Daniel Cremers. Variational uncalibrated photometric stereo under general lighting. pages 8539–8548, 2019. 1, 2

[16] H. Hayakawa. Photometric stereo under a light souce with arbitary motion. *JOSA*, 11(11):3079–3089, 1994. 2

[17] A. Hertzmann and S. Seitz. Example-based photometric stereo: shape reconstruction with general, varying brdfs. *IEEE TPAMI*, 27(8):1254–1264, 2005. 2

[18] Yannick Hold-Geoffroy, Paulo Gotardo, and Jean-François Lalonde. Single day outdoor photometric stereo. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2062–2074, 2019. 2

[19] Z. Hui and A. C. Sankaranarayanan. Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE TPAMI*, 39(10):2060–2073, 2017. 2

[20] S. Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *ECCV*, 2018. 1, 2, 4, 6, 7

[21] S. Ikehata. Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism. In *BMVC*, 2021. 1, 2, 4

[22] S. Ikehata. Universal photometric stereo network using global lighting contexts. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[23] S. Ikehata and K. Aizawa. Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *CVPR*, 2014. 2, 6, 7

[24] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Robust photometric stereo using sparse regression. In *CVPR*, 2012. 1, 2

[25] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Photometric stereo using sparse bayesian regression for general diffuse surfaces. *IEEE TPAMI*, 36(9):1816–1831, 2014. 1, 2, 4, 6, 7

[26] Yakun Ju, Junyu Dong, and Sheng Chen. Recovering surface normal and arbitrary images: A dual regression network for photometric stereo. *IEEE Transactions on Image Processing*, 30:3676–3690, 2021. 2

[27] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncalibrated neural inverse rendering for photometric stereo of general surfaces. pages 3804–3814, 2021. 1, 2, 6, 7

[28] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, pages 3744–3753, 2019. 5

[29] Junxuan Li and Hongdong Li. Neural reflectance for shape recovery with shadow handling. In *CVPR*, 2022. 2, 3, 6, 7

[30] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *ECCV*, 2022. 3, 6, 7

[31] Huiyu Liu, Yunhui Yan, Kechen Song, and Han Yu. Sps-net: Self-attention photometric stereo network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. 2

[32] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 3, 6

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 5, 6

[34] Zhuang Liu, Hanzi Mao, and Christoph Feichtenhofer Chao-Yuan Wu. A convnet for the 2020s. In *CVPR*, 2022. 3, 6

[35] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla. Px-net: Simple and efficient pixel-wise

training of photometric stereo networks. In *CVPR*, pages 12757–12766, 2021. 2, 4, 6, 7

[36] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *CVPR*, pages 1490–1497, 2013. 2

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 3

[38] Zhipeng Mo, Boxin Shi, Feng Lu, Sai-Kit Yeung, and Yasuyuki Matsushita. Uncalibrated photometric stereo under natural illumination. pages 2936–2945. IEEE Computer Society, 2018. 1, 2

[39] Y. Mukaigawa, Y. Ishii, and T. Shakunaga. Analysis of photometric factors based on photometric linearization. *JOSA*, 24(10):3326–3334, 2007. 2

[40] J. Munkberg, J. Hasselgren, T. Shen, J. Gao, W. Chen, A. Evans, T. Mueller, and S. Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *CVPR*, 2022. 6

[41] Ruth Onn and Alfred Bruckstein. Integrability disambiguates surface recovery in two-image photometric stereo. *International Journal of Computer Vision*, 5(1):105–113, 1990. 4

[42] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *CVPR*, 2013. 2

[43] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita. 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. 2

[44] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan. Self-calibrating photometric stereo. In *CVPR*, 2010. 2

[45] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. A biquadratic reflectance model for radiometric image analysis. In *CVPR*, 2012. 2

[46] B. Shi, Z. Wu, Z. Mo, D.Duan, S-K. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *CVPR*, 2016. 2, 6, 7, 8

[47] W. M. Silver. *Determining shape and reflectance using multiple images*. Master's thesis, MIT, 1980. 2

[48] P. Tan, L. Quan, and T. Zickler. The geometry of reflectance symmetries. *IEEE TPAMI*, 33(12):2506–2520, 2011. 4

[49] T. Taniai and T. Maehara. Neural Inverse Rendering for General Reflectance Photometric Stereo. In *ICML*, 2018. 2, 3

[50] A. Tiwari and S. Raman. Deepps2: Revisiting photometric stereo using two differently illuminated images. *ECCV*, 2022. 2, 6, 7

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3, 4, 5

[52] P. Woodham. Photometric method for determining surface orientation from multiple images. *Opt. Engg*, 19(1):139–144, 1980. 1, 2, 6, 7

[53] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ACCV*, 2010. 2

[54] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *NeurIPS*, 2022. 4

[55] Z. Wu and P. Tan. Calibrating photometric stereo by holistic reflectance symmetry analysis. *CVPR*, 2013. 2

[56] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 3

[57] Z Xu, K Sunkavalli, S Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. 2018. 7

[58] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi. Gps-net: Graph-based photometric stereo network. *NeurIPS*, 2020. 2, 4

[59] C. Yu, Y. Seo, and S. Lee. Photometric stereo from maximum feasible lambertian reflections. In *ECCV*, 2010. 2

[60] Yuanqing Zhang, Jiaming Sun, Xinqyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. *CVPR*, 2022. 6

[61] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pages 16259–16268, 2021. 4