Universal Photometric Stereo Network using Global Lighting Contexts

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

Figure 1. Given images of an object under different lighting conditions, our method can recover the detailed surface normal map. Images were captured by a smartphone camera with moving handheld lights under the indoor natural illumination. Please see details in Section 5.3.

Abstract

This paper tackles a new photometric stereo task, named universal photometric stereo. Unlike existing tasks that assumed specific physical lighting models; hence, drastically limited their usability, a solution algorithm of this task is supposed to work for objects with diverse shapes and materials under arbitrary lighting variations without assuming any specific models. To solve this extremely challenging task, we present a purely data-driven method, which eliminates the prior assumption of lighting by replacing the recovery of physical lighting parameters with the extraction of the generic lighting representation, named global lighting contexts. We use them like lighting parameters in a calibrated photometric stereo network to recover surface normal vectors pixelwisely. To adapt our network to a wide variety of shapes, materials and lightings, it is trained on a new synthetic dataset which simulates the appearance of objects in the wild. Our method is compared with other state-of-the-art uncalibrated photometric stereo methods on our test data to demonstrate the significance of our method.

1. Introduction

Photometric stereo is a problem of recovering the surface normal map from appearances of an object under varying lighting conditions. For decades, a broad spectrum of techniques have been proposed to expand the scope of target geometry, material and acquisition setup [15, 23, 24, 42] in the framework of the physics-based inverse rendering. Recently, advances in deep learning have been eliminating the dependence on physics-based modeling from photometric stereo methods, which contributes to handle complex optical phenomena which are hardly described in a mathematically tractable form [12, 21, 39, 44].

However, despite the long journey in this research field, each photometric stereo algorithm is still limited to a specific physical lighting model, which severely compromises its usability. In reality, most recent (semi-)calibrated [11, 21, 39, 40, 44] and uncalibrated [12, 27] photometric stereo methods still assume the single lighting in a dark environment. Others address natural lighting conditions, however both calibrated and uncalibrated methods still assume convex Lambertian surfaces and their lighting models are limited to spherical harmonics lighting [8, 17], dominant sun lighting [6, 20] and equivalent directional lighting [16, 36] which cannot represent the complex illumination.

To address this limitation, this paper proposes the “third” task in photometric stereo problem following calibrated and uncalibrated tasks. We name it universal photometric stereo (UniPS) which denotes the setup without prior assumption of physical lighting models; hence, arbitrary lighting conditions should be considered unlike calibrated and uncalibrated tasks that consider specific ones as in Table 1.1

1It would be ideal if the task were universal on materials as well. However, since some objects such as mirrors and transparent objects must be excluded, we consider only lighting conditions to be universal in this task.
In this paper, we present a first viable method for UniPS based on key insights as follows. Conventional uncalibrated photometric stereo algorithms recovered physical lighting parameters and surface normals sequentially or simultaneously; thus, constrained by specific lighting models. However, we show that the recovery of physical lighting parameters is not essential in a UniPS network, and can be replaced with the extraction of global lighting contexts from individual images by their interaction with others. In order to keep the receptive field of our network constant, contexts are extracted at the predefined canonical resolution which is independent of the input image resolution. By using global lighting contexts like calibrated lighting parameters, surface normal vectors can be recovered in similar to existing pixel-wise calibrated photometric stereo networks (e.g. [21, 22]) which easily scale to high-resolution images. While our network drops a prior assumption of lighting, its adaptation to diverse shapes, materials and lightings has to be ensured by training data. Since existing training datasets for photometric stereo networks are limited to the single, directional lighting setup, we create a dataset for our task by physically rendering the appearance of objects with more than 10,000 combinations of shape, material and lighting using high quality commercial 3-D assets. We also create an evaluation dataset with 50 sets of attributes using different assets to compare our method with state-of-the-art uncalibrated photometric stereo algorithms specifically designed for directional lighting [12] and natural lighting [17, 36]. Finally, the qualitative evaluation demonstrates that our method even works for real objects under the challenging spatially-varying lighting conditions that were conventionally considered to be intractable (See Fig. 1).

2. Related Work

Here we briefly review uncalibrated photometric stereo methods under the directional or natural lighting. Due to the space limit, other categories (e.g. calibrated setup [21–25], near-light setup [26,40] or multi-view setup [29,38]) are not included in this survey. However, we emphasize that all the methods more or less assume any of physical lighting models; hence, are limited to specific lighting setups.

Uncalibrated, Directional Lighting: Since Woodham [46] presented the first Lambertian photometric stereo algorithm in 1980, methods following its setup have been called calibrated photometric stereo which assumes an orthographic camera and a known single directional lighting. The uncalibrated task is almost identical to the calibrated one except that lighting parameters are unknown. Until very recently, most uncalibrated photometric stereo algorithms assumed Lambertian integrable surfaces and the goal was to resolve the General Bas-Relief ambiguity [18] between geometry and light. Various cues were employed for resolving this ambiguity, which include inter-reflections [13], entropy of surface albedo [7], color profile [41], diffuse maxima [14], reflectance symmetry [47] and perspective geometry [37]. Though there were very few methods for non-Lambertian surfaces due to their ill-posed nature, Lu et al. [35] utilized the statistical distribution of the intensity profile to recover isotropic non-Lambertian surfaces.

In 2018, the first deep neural network for uncalibrated photometric stereo was presented [11]. However, it simply dropped the lighting channel from its calibrated variant presented at the same time and the performance was quite limited. Hence, authors extended their work to the two-step approach [10] where only lighting information was firstly recovered, and then it was used as input of the calibrated photometric stereo network. Later, authors further updated their work by feeding the surface normal estimation result back to the lighting prediction to improve its accuracy [12]. Building upon this work, Kaya et al. [27] have recently utilized the lighting prediction result from [12] to recover surface normals in the neural inverse rendering framework.

Uncalibrated, Natural Lighting: Although most photometric stereo methods assume the single directional lighting condition, some literature exists which addresses natural lighting conditions. However, inversely decomposing the natural light reflected on non-Lambertian surfaces is intractable in the inverse rendering framework, so it is common to assume convex Lambertian surfaces and approximated lighting models [8,9,16,17,36].

The first uncalibrated photometric stereo algorithm under natural lighting was proposed by Barsri et al. [8] which approximated natural lighting as global first-order spherical harmonics. Though a global concave-convex ambiguity exists in estimated surface normals due to the lighting approximation with an orthographic camera, Brahimi et al. [9] recently proved that the perspective integrability constraint makes the problem well-posed. Mo et al. [36] presented another uncalibrated photometric stereo algorithm under natural lighting. They proposed the equivalent directional lighting model which decomposed the entire task into the patchwise directional uncalibrated photometric stereo problem. This work was later extended by authors where MRF-based global optimization and rotation averaging were introduced for a better normal patch integration [16]. These methods worked under an orthographic camera, however when solving the global orthogonal ambiguity with integrability, there was a binary ambiguity left which requires to be solved manually and a non-integrable surface can not be recov-
Since the ultimate goal of UniPS is to realize a truly practical photometric stereo method, the available geometries and materials (except for extreme cases such as transparent and perfect mirror objects) should be diverse as well and the number of input images and their resolution should be arbitrary. Although accepting arbitrary number of input images was considered as an important requirement in recent works (e.g. [11, 21, 44]), there are still some methods that do not meet this requirement (e.g. [31]). Scalability is basically not a problem in pixel-by-pixel algorithms (e.g. [21, 24]), but it often becomes a major issue in methods that use the entire image information such as convolutional neural networks and global optimization (e.g. [12, 36]).

### 4. Method

#### 4.1. UniPS Network with Global Lighting Contexts

This paper presents a first viable universal photometric stereo network. As illustrated in Fig. 2, our network consists of an encoder and decoder. The encoder extracts *global lighting contexts* from images and an object mask, which is a generic lighting representation that corresponds to physical lighting parameters (e.g. light direction) in deep uncalibrated photometric stereo networks [10, 12]. The decoder takes all raw image values and the interpolated global lighting context at each pixel and predict its surface normal.

Our network architecture has two major differences from basic encoder-decoder architectures that take a single image as input. First, our network takes multiple images as input; hence, features must be embedded in the latent space considering the interaction of them. Therefore, we perform feature communication in the encoder and aggregation in the decoder, respectively. Second, unlike typical architectures where encoded features are directly passed to the decoder, we use different working resolutions for the encoder and decoder. The working resolution for the decoder is same as the original image resolution, but the encoder takes as input images that have been resized to the pre-defined **canonical resolution**, which is basically smaller than the original resolution and its output is passed to the decoder after inversely converted for the decoder’s working resolution.

There are two major advantages to use different working resolutions. First, the scalability to the image size is ensured because memory requirements for the encoder depend only on the canonical resolution, not on the original one, while the decoder processes each pixel one by one. The second and more important reason is to keep the receptive field of the encoder invariant to the input image size. Without this, the networks’ receptive field may not cover the entire object in extremely high-resolution test images.

#### 4.2. Framework Components

**Preprocessing:** The pixel value range of individual images can significantly vary under different lightings. There-
Preprocessed crops of the images and mask are bilinearly resized to the predefined $s \times s$ canonical resolution. They are then concatenated $C_{\{1,...,q\}} \in \mathbb{R}^{s \times s \times (3+1)}$ and passed to the encoder. The output of the encoder is embedded features for individual images whose size is a quarter of the canonical resolution $G_{\{1,...,q\}} \in \mathbb{R}^{\frac{s}{4} \times \frac{s}{4} \times d_e}$ where $d_e$ is the embed dimension. We call them as global lighting contexts because the difference of features among images should only be attributed to the difference of lighting conditions. Global lighting contexts are also an analogy of physical lighting parameters similarly recovered by the conventional uncalibrated photometric stereo methods. However, unlike physical parameters were basically assigned to each image globally (e.g. light direction [10], spherical harmonics [17]), a unique global lighting context is assigned at each location (therefore we name context’s'); hence has the capacity to represent spatially-varying phenomena such as near-lighting, inter-reflections and cast shadows.

The encoder mutually embeds global lighting contexts via the image-wise feature extraction and inter-image feature communication. The former extracts multi-scale feature maps using a backbone (e.g. SwinTransformer [32]) followed by the basic multi-scale feature fusion [48]. The latter propagates features across images without changing the feature dimension by pixelwisely employing a single Transformer layer [45] as with the recent calibrated photometric stereo network [22]. As illustrated in Fig. 3, there are multiple feasible designs of the encoder depending on how information propagate via inter-/intra-image interactions. The best design of them will be discussed in the ablation study in Section 5.1. As it will turn out, the feature communication should be conducted just before the multi-scale feature fusion (i.e. pre-fusion). Please refer to the supplementary for more details about individual implementations of the feature extraction and communication.

**Decoder:** For each coordinates $x$ of the original resolution, the decoder inputs a set of image values $I_{\{1,...,q\}}(x)$ and a global lighting context $G_{\{1,...,q\}}(S(x))$, and outputs the unit normal vector at the location. Here, $S$ is the sampling operation to fetch the value of the global lighting context corresponding to $x$ using bilinear interpolation. Note that avoiding the resize of global lighting contexts to the original resolution also contributes to the scalability.

The decoder is composed of the feature aggregation and the surface normal recovery. The former further propagates information across different lighting conditions with extra Transformer layers, and then squeezes the lighting channel with the pooling by multi-head attention (PMA) [28]. The output of PMA is a pixelwise feature vector independent of number of input images. The latter feeds the feature to the multi-layer perceptron with one hidden layer to output the surface normal vector. We should note that when no communication is conducted in the encoder as in Fig. 3, it can be deemed that global lighting contexts have been extracted at the aggregation step in the decoder.

**Training Loss:** The mean squared error between ground truth and predicted surface normals is used as the training loss. Because of the efficiency, we compute the loss only on samples from specific locations, not from the entire image. For each object, we uniformly sample pixels at the original resolution whose centers can be projected exactly onto centers of pixels at the canonical resolution. Then, we further sample pixels from a limited number of random locations (e.g. 2500) at the original resolution which allows for the normal estimation with a sub-pixel global lighting context.

**Network Analysis:** Our method is data-driven; hence satisfies the requirements of UniPS. This is achieved by global lighting contexts, which eliminate the need for the physical lighting parameter recovery and makes our method feasible to handle complex spatially-varying lighting effects.

Our framework is scalable by introducing the canonical resolution for encoding global lighting contexts and pixelwisely applying the decoder. The network’s receptive field is independent of the input image size as well as there is no upper limit to the test image size. By utilizing the self-attention mechanism for the feature interaction across images, the network is applicable to an arbitrary number of images as long as the computational resource permits. Given this capacity, the availability of varying shapes, materials and lightings are secured by training data.

### 4.3. PS-Wild Training Dataset

We need sufficient training examples that properly characterize the universal photometric stereo task. Though two major synthetic photometric stereo datasets (i.e. Blobby and Sculpture [11], CyclesPS [21]) have been presented to date, they were for the methods under the single directional lighting and not applicable to UniPS. Therefore, we present a new photometric stereo dataset, named PS-Wild, which sim-
ulates the appearance of objects with diverse geometries and materials captured under various lighting conditions in the wild. The idea is simply to ask a physically-based renderer (i.e., Blender Cycles [2]) to synthesize a large number of images under general lightings, taking full advantage of high quality commercial 3-D assets.

To assemble an appropriate collection of object appearances for our PS-Wild dataset, we browsed through online marketplaces looking for 3-D assets that satisfied three main desiderata. First, we should use 3-D models that have sufficient complexity. We don’t want the objects whose surface normal distribution deviates from ones in the wild (e.g., too many planar objects or low-poly models are not adequate). Second, we want 3-D materials to be as diverse as possible. In our problem, we don’t assume specific materials as well as lighting models, so the training data needs to be chosen to cover diverse materials in the wild. In addition, textures need to be realistic. In the existing synthetic datasets, the entire image was rendered with a single BRDF [11], or pixelwisely different random BRDFs [21, 33], but the actual surface texture has some realistic rules as mentioned in previous studies [7, 15]. Hence, it is worth using texture maps that were designed by professionals. Third, we want to render scenes under diverse lighting conditions. In this regard, the most practical method we found is to use the HDR image-based (HDRI) lighting that covers various indoor and outdoor scenes. To add more details, this method uses an HDR environment map to place an omnidirectional light source. Each ray from the light is self-shielded before it hits the surface, so a variety of spatially-varying lighting effects (e.g., cast shadows, near-lighting) are baked.

Based on these desiderata, we choose Adobe Stock [1] as our assets to create the training dataset. This collection consists of over 17,000 3-D assets including 3-D models, materials (texture maps) and lightings (environment maps). From all available assets, we downloaded 410 3-D models, 926 materials and 31 lightings. Most of data in Adobe Stock assets are 3-D models and we actually used all the materials and lightings available at the time we accessed there (Oct. 2021). The number of assets itself may not seem very large, but in reality, a countless number of images can be synthesized by various augmentation techniques such as rotation and color transformation.

After acquiring 3-D assets, we apply our computational pipeline to generate images with ground truth surface normal maps. For each 3-D object in assets, we randomly assign one material and one environment map. Then we apply several random rotations of the object until the Shannon entropy of its surface normal distribution becomes more than a threshold (i.e., 4.0 in our method) and if the maximum entropy is less than the threshold, the 3-D model is discarded from assets. We then scale the object to make sure that the object’s outline reaches the edge of the image. We render 10 of $512 \times 512$, 16-bit images per object. For each rendering, we randomly rotate the environment map on the spherical axis to make the variation of lighting conditions and automatically adjust the exposure of the camera to make the dynamic range of rendered images consistent. We turn on the ray tracer to render the cast shadow and inter-reflection for adding the spatially-varying lighting effects. Finally, we got 10,099 objects with a different pose, material and lighting.

4.4. PS-Wild Test Dataset

We also create a test dataset for the evaluation purpose. The computational pipeline of generating images is same as one for the training dataset but different 3-D assets are used for the fair evaluation; 25 objects from CGTrader [3], 50 materials from ShareTextures [4] and 50 environment maps from sBL Archive [5]. For each 3-D model, we assign two sets of a material and environment map resulting in 50 different sets of the object, material and environment map. Unlike training dataset, we carefully pick six textures per each texture category; Concrete, Fabric, Floor, Ground, Wood and Metal as categorized in ShareTextures. In order to properly evaluate the performance of a method under various lighting conditions, we render images for the same set of the object and material using three different lighting methods; (a) single directional lighting (uniformly sampled), (b) HDRI lighting (same as training) and (c) mixture of (a) and (b). The image resolution is also $512 \times 512$ but the number of images is 32 to evaluate the performance with varying number of input images.

**Dataset Analysis:** Our test dataset characterizes the universal photometric stereo task. Objects with both convex and non-convex geometries and a variety of spatially-varying materials involving diffuse, specular and metallic are rendered with three different lighting methods including the HDRI lighting which exhibits challenging spatially-varying lighting effects. Therefore, methods that assume a specific physical lighting model are not suitable for this dataset.

5. Results

We conduct experiments on synthetic and real test data. We first ablate the important design elements of our architecture and then we compare the proposed universal photometric stereo network with the previous state-of-the-arts on the uncalibrated photometric stereo task [10, 17, 36]. For the convenience, we henceforth refer to the proposed method as UniPS-GLC (Universal Photometric Stereo network using Global Lighting Contexts) as necessary.

**Training Details:** Our network was trained from scratch on a NVIDIA Quadro RTX 8000 machine with AdamW [34] optimizer for 20 epochs using a step decay learning rate schedule ($\times 0.8$ every three epochs). A batch size of 3, an initial learning rate of 0.0001, and a weight decay of 0.05
Table 2. Comparison with different image feature extractors.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional</td>
<td>19.7</td>
<td>29.5</td>
<td>48.6</td>
<td>50.6</td>
</tr>
<tr>
<td>HDRI</td>
<td>16.8</td>
<td>23.2</td>
<td>45.8</td>
<td>44.0</td>
</tr>
<tr>
<td>Dir.+HDRI</td>
<td>16.1</td>
<td>24.0</td>
<td>45.2</td>
<td>45.0</td>
</tr>
</tbody>
</table>

Table 4. Comparison with different canonical resolutions.

<table>
<thead>
<tr>
<th></th>
<th>128 × 128</th>
<th>256 × 256</th>
<th>512 × 512</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional</td>
<td>19.9</td>
<td>19.7</td>
<td>N.A</td>
</tr>
<tr>
<td>HDRI</td>
<td>18.1</td>
<td>16.8</td>
<td>N.A</td>
</tr>
<tr>
<td>Dir.+HDRI</td>
<td>17.0</td>
<td>16.1</td>
<td>N.A</td>
</tr>
</tbody>
</table>

Table 6. Comparison with different feature aggregation methods.

<table>
<thead>
<tr>
<th></th>
<th>Max-pool</th>
<th>TF+PMA (3L)</th>
<th>TF+PMA (6L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional</td>
<td>59.3</td>
<td>19.7</td>
<td>37.5</td>
</tr>
<tr>
<td>HDRI</td>
<td>39.3</td>
<td>16.8</td>
<td>29.4</td>
</tr>
<tr>
<td>Dir.+HDRI</td>
<td>39.0</td>
<td>16.1</td>
<td>30.6</td>
</tr>
</tbody>
</table>

were used. The number of random samples from the original resolution was fixed by 2500. In total, it took roughly 48 hours to train the network for each configuration.

**Inference Time:** The inference time of our method depends on the number and resolution of input images. In the case of 32 of 512 × 512 images as input, it takes less than a few seconds excluding IO on GPU. This is slightly slower than the simpler deep photometric stereo networks (e.g. 0.5 sec in [10]), however more than hundred times efficient than inverse rendering based methods [17, 27, 36, 44].

**Evaluation metric:** Evaluation is based on the mean angular errors (MAE) between predicted and true surface normal maps measured in degrees (0 to 180). In our evaluation, we apply each algorithm to PS-Wild test dataset with three different lighting methods and discuss the result mainly based on the averaged MAE over 50 different objects.

### 5.1. Ablation study

**Base architecture:** As needed, we define the base architecture for the ablation study with following design elements: SwinTransformer [32] for feature extraction; No communication in the encoder; 256 × 256 canonical resolution; Three stacks of transformer layers followed by PMA for the feature aggregation in the decoder; In ablation, only the target property was changed from this base architecture.

**Encoder:** Table 2 shows the comparison of four different image feature extractors: SwinTransformer [32], Vision-Longformer [49] and ResNet-50/101 [19]. We observe that Transformer-based encoders, especially SwinTransformer outperformed ResNets probably due to the larger receptive field of the Transformer model. Because the main purpose of this work is to present a viable method to demonstrate our ideas, the further discussion remains future work.

**Decoder:** Table 3 compared four different variants of the feature aggregation; no communication in the encoder (Base) during the image feature extraction (Dur-Ext), before the multi-scale fusion (Pre-Fus) and after the multi-scale fusion (Post-Fus) as illustrated in Fig. 3. As have been mentioned already, the best performance was obtained when the communication was done just before the multi-scale fusion. Interestingly, the second best was obtained when no communication among images was done in the encoder, but only in the decoder. We analysed this result and found that the communication during the image feature extraction simply broke the feature extraction process and the post-fusion communication made the optimization unstable since it is equivalent to increasing the transformer layers in the decoding process which is consistent with later results.

Table 4 compared different canonical resolutions either from 128×128 or 256×256 and we observe that 256×256 worked slightly better. This result indicates that though the lower resolution is helpful to see the entire object, important information for details could be discarded at very low resolution. On the other hand, it is surprising to see that 32×32 global lighting contexts from 128×128 canonical resolution still provides a reasonable reconstruction for 512×512 output resolution. Unfortunately, we couldn’t get the result of 512×512 canonical resolution due to the memory limit.

Finally, we compared our spatially-varying global lighting contexts with the spatially-uniform ones by applying the global average pooling to shrink \( \{ G \} \in \mathbb{R}^{s \times 1 \times d_c} \) to \( \{ G_{uni} \} \in \mathbb{R}^{1 \times 1 \times d_c} \) and feeding the spatially same vector to the decoder. In reality, this is a similar procedure in existing uncalibrated photometric stereo networks [10, 27] where the single lighting parameter is firstly recovered for each image and used as the input of the surface normal predictor. However, Table 5 shows that the network with the uniform context always provided the trivial solution since it couldn’t capture spatially-varying lighting effects.

**Decoder:** We compared three different variants of the pixel-wise feature aggregation strategy from a max pooling, three Transformer layers with PMA and six Transformer layers with PMA; The further to the right we go, the more complex interaction can theoretically take place. Note that the latent vector size in the Transformer layer was fixed by 384 and
Different number of input images:

Table 7. Different number of input images (Ours, pre-fusion).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional</td>
<td>35.4</td>
<td>22.6</td>
<td>18.8</td>
<td>17.5</td>
<td>17.0</td>
</tr>
<tr>
<td>HDRI</td>
<td>27.2</td>
<td>19.5</td>
<td>16.5</td>
<td>14.7</td>
<td>14.5</td>
</tr>
<tr>
<td>Dir.+HDRI</td>
<td>28.5</td>
<td>20.6</td>
<td>17.4</td>
<td>15.6</td>
<td>13.8</td>
</tr>
</tbody>
</table>

Comparison with other methods (pre-fusion):

Table 8. Comparison with other methods (pre-fusion).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional</td>
<td>17.0</td>
<td>17.7</td>
<td>32.9</td>
<td>33.0</td>
</tr>
<tr>
<td>HDRI</td>
<td>14.5</td>
<td>24.8</td>
<td>30.4</td>
<td>37.5</td>
</tr>
<tr>
<td>Dir.+HDRI</td>
<td>13.8</td>
<td>31.5</td>
<td>30.8</td>
<td>32.5</td>
</tr>
</tbody>
</table>

The hidden dimension of the feedforward layer was 1024. The result is illustrated in Table 6. Though there is no surprise to see the simple max-pooling didn’t work at all, the reason for the worse results for the deeper Transformer networks seemed to be due to the over-fitting. It should be noted that the Transformer network is known to be a difficult network to train, and careful adjustment of the hyperparameters may yield different results.

5.2. Quantitative Evaluation on PS-Wild Test Data

We compared our method to state-of-the-arts on the uncalibrated photometric stereo problem on our synthetic test dataset. The first comparison point is GCNet [10] which is a learning-based method under the directional lighting condition that alternately estimates physical lighting parameters and surface normals. The second point is the matrix-based patch merging (MPM) [36] which is the Lambertian photometric stereo method for uncalibrated natural lighting based on the equivalent directional lighting model. The third point is the variational uncalibrated photometric stereo method (Variational) [17] which approximates Lambertian reflectance model through a spherical harmonic expansion. We used authors’ official codes for the evaluation and our GLC-UniPS architecture was “base + pre-fusion” configuration that has yielded the best result so far.

The results are illustrated in Fig. 4 and Table 8. We observe that our UniPS-GLC reasonably worked independent of the lighting methods. As expected, GCNet [10] worked fine for objects under the directional lighting condition (e.g. ID 48), however problematic with non-uniform materials and non-directional lightings. Overall, MPM [36] and Variational [17] were inferior to GCNet for diverse objects due to their assumptions of convex, Lambertian surfaces. Especially, MPM had a problem in handling spatially varying non-Lambertian textures or non-convex geometries (e.g. ID 37) and variational also had a problem in handling non-Lambertian reflections especially for metallic surfaces (e.g. ID 31). Interestingly, better performance of our method was observed for objects of the same shapes and materials under the environment lighting than the directional one. This result may be attributed to the fact that in an image under the directional lighting condition, the shadow or inter-reflection behaves negatively, whereas under the environment light, they positively yields spatially-varying lighting effects as a source of information.

The input images and recovered normal map for individual objects are presented in the supplementary.
5.3. Qualitative Evaluation on Real Data

Fig. 1 validates our method (base+pre-fusion) on real images. Objects of varying materials such as ceramic, metallic and clear coat were captured by a 8-bit smartphone camera under a few hand-held area lights and/or the indoor natural illumination. The area light sources were intentionally put within 30 cm of an object so that strong spatially-varying lighting effects were observed. Although quantitative evaluation is not possible due to the absence of the ground truth, we observe that overall reasonable normal maps were recovered even under these extremely challenging setups. We also observe that the very low resolution of GLiCo (i.e., 64 × 64) is sufficient to recover surface details.

For a more objective evaluation, Fig. 5 provides a comparison with GCNet, MPM and Variational. Images were captured by putting a single area light near an object under the static indoor natural illumination; therefore, could not be represented by physical lighting models such as the directional lighting nor spherical harmonics. We picked four objects with different difficulties and included two “spherical” objects (i.e., Grapes and Metallic Painted Spheres) so that the quality of normal maps could be evaluated based on their spherical parts. Grapes exhibit complex cast shadows because of its non-convex geometry. Metallic Painted Spheres consists of four metallic-painted balls with different surface roughness and structures which exhibit strong inter-reflections with each other. Fortune Cat contains large “black” regions, which are known to be difficult to distinguish from shadows. Bowl includes multiple challenging objects put in a bowl which is hard to resolve the convex-concave ambiguity. We emphasize that these are extremely difficult objects that existing photometric stereo methods have not even attempted to handle them. Though our results are still far from perfect (i.e., we clearly observe errors in our results due to inter-reflection in Metallic Painted Spheres and convex geometries in Bowl), our method obviously recovered most reasonable surface normal maps.

6. Conclusion

This paper tackled a new photometric stereo task, named universal photometric stereo (UniPS) which drops prior assumptions of physical lighting models. To this end, we presented the first viable UniPS network based on the generic lighting representation named global lighting contexts. We also presented synthetic training and evaluation datasets for our UniPS task and our extensive evaluation demonstrated the performance of our method.

Despite the significant progress toward a practical photometric stereo method, there are some limitations remained. First, the deeper analysis of global lighting contexts is missing. To explore its potential for applications other than the surface normal recovery is a future work. Second, our training method specifies a camera model (e.g., orthographic, linear camera) and we may also want to remove this dependence. Finally, we need a real dataset with the ground truth for evaluating UniPS methods in a more quantitative manner.
References

[17] 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, 3, 4, 5, 6, 7
[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2019. 5


