Revisiting One-stage Deep Uncalibrated Photometric Stereo via Fourier Embedding

Yakun Ju, *Member, IEEE,* Boxin Shi, *Senior Member, IEEE,* Bihan Wen, *Senior Member, IEEE,* Kin-Man Lam, *Senior Member, IEEE,* Xudong Jiang, *Fellow, IEEE,* Alex C. Kot, *Life Fellow, IEEE*

Abstract—This paper introduces a one-stage deep uncalibrated photometric stereo (UPS) network, namely Fourier Uncalibrated Photometric Stereo Network (FUPS-Net), for non-Lambertian objects under unknown light directions. It departs from traditional two-stage methods that first explicitly learn lighting information and then estimate surface normals. Two-stage methods were deployed because the interplay of lighting with shading cues presents challenges for directly estimating surface normals without explicit lighting information. However, these two-stage networks are disjointed and separately trained so that the error in explicit light calibration will propagate to the second stage and cannot be eliminated. In contrast, the proposed FUPS-Net utilizes an embedded Fourier transform network to implicitly learn lighting features by decomposing inputs, rather than employing a disjointed light estimation network. Our approach is motivated from observations in the Fourier domain of photometric stereo images: lighting information is mainly encoded in amplitudes, while geometry information is mainly associated with phases. Leveraging this property, our method "decomposes" geometry and lighting in the Fourier domain as guidance, via the proposed Fourier Embedding Extraction (FEE) block and Fourier Embedding Aggregation (FEA) block, which generate lighting and geometry features for the FUPS-Net to implicitly resolve the geometry-lighting ambiguity. Furthermore, we propose a Frequency-Spatial Weighted (FSW) block that assigns weights to combine features extracted from the frequency domain and those from the spatial domain for enhancing surface reconstructions. FUPS-Net overcomes the limitations of two-stage UPS methods, offering better training stability, a concise end-to-end structure, and avoiding accumulated errors in disjointed networks. Experimental results on synthetic and real datasets demonstrate the superior performance of our approach, and its simpler training setup, potentially paving the way for a new strategy in deep learning-based UPS methods.

Index Terms—3D reconstruction, photometric stereo, Fourier transform

1 Introduction

Photometric stereo (PS) aims to recover the surface normal of an object from diverse shading cues in multiple images with different lighting conditions [1]. Compared with geometric stereo methods, photometric stereo methods can capture pixel-wise high-frequency details on textureless surfaces. Therefore, PS plays a crucial role in recovering fine-detailed surfaces, particularly in scientific and engineering fields like cultural relics digitization [2], forensics [3], and industrial detection [4].

Classic PS [1] assumes that only Lambertian (diffuse) reflectance exists on the surface of the target object. However, real-world objects rarely exhibit pure Lambertian reflectance, which impacts the linearly proportional relationship between images and surface normals. Previous methods use complex reflectance modeling [5], [6], outlier rejection [7], [8], or deep learning-based networks [9], [10] to handle the recovery of normals from non-Lambertian surfaces. Mathematically, we express the non-Lambertian property via the bidirectional reflectance distribution function (BRDF), depending on the material of the object. In this case, the relationship between a measured pixel intensity o and the corresponding surface point with normal $\boldsymbol{n} \in \mathbb{R}^3$ being illuminated by lighting with direction $\boldsymbol{l} \in \mathbb{R}^3$ and intensity $e \in \mathbb{R}$, observing from view direction $\boldsymbol{v} \in \mathbb{R}^3$, *i.e.*, the image formation model, can be expressed as

$$o = e\boldsymbol{\rho}(\boldsymbol{n}, \boldsymbol{v}, \boldsymbol{l}) \max(\boldsymbol{n}^\top \boldsymbol{l}, 0) + \epsilon, \tag{1}$$

where $\boldsymbol{\rho}$ stands for the BRDF and ϵ represents the global illumination noise in images, such as cast shadows and inter-reflections.

Calibrated photometric stereo (CPS) methods [12], [13] rely on knowledge of the lighting direction ($\boldsymbol{l}$) for each image. However, calibrating the lighting directions involves complex operations and relies on specialized instruments, making it impractical for real-world applications. Conversely, uncalibrated photometric stereo (UPS) [14], [15] can estimate surface normals without lighting information, which however faces challenges in resolving the geometry-lighting ambiguity, such as the Generalized Bas-Relief (GBR) ambiguity [16]. Unfortunately, resolving the geometry-lighting ambiguity usually requires the assumption of a simplified Lambertian reflectance model [17], [18]. Although some methods [15], [19] can handle surfaces with general BRDFs, they are restricted to a uniform distribution of lighting directions.

- *Yakun Ju is with the School of Computing and Mathematical Sciences, University of Leicester, Leicester, United Kingdom (e-mail: kelvin.yakun.ju@gmail.com).*
- *Yakun Ju, Bihan Wen, Xudong Jiang, and Alex C. Kot are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: kelvin.yakun.ju@gmail.com, bihan.wen@ntu.edu.sg, exdjiang@ntu.edu.sg, eackot@ntu.edu.sg).*
- *Boxin Shi is with the State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China (e-mail: shiboxin@pku.edu.cn).*
- *Kin-Man Lam is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: enkmlam@polyu.edu.hk).*
- *Corresponding Author: Boxin Shi*

Fig. 1. Motivation: We observed that shape information and lighting information can be "decomposed" in the Fourier domain. Amplitude and phase are generated by the Discrete Fourier Transform (DFT) and the compositional images are obtained by Inverse DFT (IDFT) [11]. (a) We swapped the phase components of two photometric stereo images of different lighting directions of the same object "Buddha". The compositional images maintain similar illuminations as the original two. (b) We further swapped the phase components of two photometric stereo images with different objects "Reading" and "Goblet". The results produce two compositional images with exchanged objects.

Recently, deep learning-based UPS methods have demonstrated impressive results in handling general reflectance surfaces without additional clues, owing to the powerful capabilities of deep neural networks [20], [21], [22]. UPS-FCN [20] is a representative method capable of addressing the UPS problem without the explicit need to learn lighting directions. However, the performance of UPS-FCN falls short of expectations due to the complex coupling among shading cues, encompassing unknown lighting directions, surface normals, and reflectance properties, *i.e.*, the geometry-lighting ambiguity (GBR ambiguity [16]). Therefore, almost all subsequent deep learning-based UPS methods [21], [22], [23], [24] adopt an explicit light estimation strategy, which first estimates the lighting directions and then maps the surface normals using both the estimated lighting information and input images, namely two-stage methods.

However, the two-stage network strategy brings some other challenges. First, existing methods [21], [22], [24] concatenate the expanded lighting directions with the input images and use CNN-based encoders to approximately decouple the features of surface normals. Although this approach has achieved good results, the two-stage methods suffer from training instability. These methods need to separately train the light estimation network and normal estimation network. In addition, as the two-stage light calibration network and normal estimation network are disjointed, the error in explicit light calibration will propagate to the second stage and cannot be eliminated. Furthermore, current deep UPS methods [21], [22], [23], [24] have to convert the estimation of lighting direction from regression of an exact vector to classification in a discretized space. This is because classifying lighting directions into predefined bins of angles is much easier than directly regressing the unit vector itself. However, this conversion limits the learning of accurate lighting directions. Therefore, these methods may struggle to balance learning difficulty and accuracy, posing challenges for effectively estimating surface normals.

To address the aforementioned challenges, we propose a novel framework that utilizes a one-stage Fourier Embedding network to handle UPS, namely FUPS-Net, eliminating the need for explicit learning of lighting directions. Our Fourier-based approach diverges significantly from existing methods that process images in the spatial domain. Our method is motivated by our observation of photometric stereo images in the Fourier domain [11]. As shown in Fig. 1 (a), swapping the phases of two photometric stereo images under different lighting directions yields two compositional photometric stereo images with unchanged illuminated lights. In contrast, as shown in Fig. 1 (b), swapping the phase of photometric stereo images captured with different objects results in the exchanged objects. These phenomena suggest that, in photometric stereo images, lighting information mainly resides in the amplitude, while geometry information correlates with the phase (further discussion is provided in Section 3.1). In other words, the geometry and lighting can be represented by phase and amplitude in the Fourier domain. Therefore, amplitude spectrum can provide lighting features for the UPS network and implicitly solve the GBR ambiguity [16]. The observations inspire the main design of our framework, realized by the Fourier Embedding Extraction (FEE) block (Section 4.1) and Fourier Embedding Aggregation (FEA) block (Section 4.2), which process the information of lighting and shape within a Fourier-embedded one-stage network.

Furthermore, we found that the features from the Fourier domain extract spatially global information because each frequency component contains some information from the whole spatial domain (further discussion is provided in Section 3.2). Therefore, we propose a Frequency-Spatial Weighted (FSW) block to extract global information in photometric stereo images (Section 4.3). This block is crucial because long-range spatial context is essential for accurate feature extraction, particularly for capturing shadows and inter-reflections. The FSW block utilizes a Normalized High-frequency (NHF) map to calculate weights for combining the global Fourier frequency branch and the local spatial branch. The NHF map effectively provides the fusion of higher local spatial information and lower global Fourier frequency information on flat regions of the object while showing the opposite pattern in areas with cast shadows, specular highlights, and complex structures.

In fact, our approach shares some similarities with recent methods [25], [26] for the universal photometric stereo task, both of which implicitly extract lighting information from observations rather than learning specific lighting directions

and intensities (*e.g.*, environmental light, near-point light). However, Universal PS methods [25], [26] are constrained by minimal lighting variations due to the interacted lighting extraction strategy. In contrast, our method leverages strong prior knowledge from Fourier transform decomposition as guidance, offering more robust implicit features.

Our method offers several advantages. Firstly, it avoids the difficulty of explicitly learning exact lighting directions in previous UPS networks by implicitly decomposing lighting and geometry information through the Fourier transform. Secondly, we leverage the Fourier transform to capture global information of the image with low computational cost. We conduct thorough ablation experiments to demonstrate the effectiveness of the proposed blocks. Furthermore, we demonstrate the performance of our FUPS-Net in addressing the UPS problem across various benchmark datasets including DiLiGenT [27], DiLiGenT10^2 [28], synthetic test data [29], and the real photoed Light Stage Data Gallery dataset [30].

In summary, this paper focuses on establishing a data-driven one-stage UPS network using the embedded Fourier transform. Our contributions are outlined as follows:

- We first investigate the utilization of Fourier frequency information in deep photometric stereo. This approach is based on the observation that the lighting and geometry information of a photometric stereo image can be represented by phase and amplitude in the Fourier domain.
- We introduce a one-stage Fourier-embedded UPS network with FEE and FEA blocks. This network implicitly learns lighting directions within a concise end-to-end structure.
- We propose an FSW block to assign prior weights, efficiently combining Fourier frequency features and spatial features for improving surface normal recovery.
- The proposed FUPS-Net offers a simpler one-stage end-to-end training setup and faster running time, while achieving superior results and avoiding the accumulation of errors found in previous methods.

2 Related Work

2.1 Calibrated Photometric Stereo (CPS)

Classic photometric stereo [1] assumes that only Lambertian (diffuse) reflectance exists on the surface of the target object, enabling shape recovery using the least squares method. However, real-world objects seldom exhibit purely Lambertian reflectance. Traditional photometric stereo algorithms have addressed non-Lambertian photometric effects through various approaches, including BRDF modeling [5], [6], outlier region rejection [7], [8], and exemplar-based techniques [31], [32]. Readers can refer to [27] for a comprehensive survey on these non-learning-based methods.

In recent years, deep learning-based methods have been widely used in the context of photometric stereo [9], [10], [20], [33], [34], [35], [36], [37]. DPSN [37] pioneered a fully connected deep photometric stereo network for estimating pixel-wise surface normals. However, it is limited to a fixed number and sequential order of observations. To handle a variable number of observations, some works have mapped pixels into an observation map in a per-pixel manner [10], [33], [38], while others have extracted global cues from patches for estimation of normals in an all-pixel manner [9], [20], [35]. Subsequent techniques [34], [36] have combined both strategies to extract local and global features for more effective estimation of normals. Recently, some works [39], [40] have further applied the Transformer with the self-attention mechanism [41] in the context of photometric stereo, which aims to capture long-range context and facilitate the aggregation of features. For further details, surveys by [42], [43] offer insights.

However, these approaches assume known lighting conditions and cannot effectively handle uncalibrated photometric stereo. Calibrating light sources can be a tedious process, requiring professional devices and may be unavailable in real-world applications. It would be more convenient for the community if photometric stereo methods could operate without the need for ground-truth lighting directions.

2.2 Uncalibrated Photometric Stereo (UPS)

UPS methods aim to automatically calibrate lighting conditions, eliminating the need for explicit knowledge of lighting directions. However, solving UPS introduces geometry-lighting ambiguity, such as GBR ambiguity [16], which is an inherent inability due to the lack of light source directions. To address this ambiguity, traditional methods have been developed to provide additional knowledge, such as inter-reflections [44], specular spikes [45], parametric specular reflection [46], isotropic specular reflection [47], *etc*. However, these methods necessitate manual labeling of mirror-like specularities for computation [45] or assume uniformly distributed albedos [46]. These approaches either rely on unrealistic assumptions or exhibit instability in their solutions, leaving a gap in their applicability to real-world scenarios.

With the recent advancements in neural networks, deep learning-based methods have achieved state-of-the-art performance in addressing the UPS problem. These neural network methods learn prior information for solving the GBR ambiguity from a large amount of training data with ground truth. UPS-FCN [20] first addresses the UPS problem without the input of lighting directions. However, the performance of UPS-FCN is limited because it cannot solve the ambiguity without the learned lighting features. Later, Chen *et al.* [21], [24] proposed two-stage networks, which first estimate light conditions and then learn surface normals with both lighting information and images, thereby solving this ambiguity. Later advancements refined this pipeline by employing a differentiable neural architecture search (NAS) strategy to automatically discover the most efficient neural architecture [22]. Additionally, uncalibrated neural inverse rendering approaches were utilized to handle unknown lighting conditions [23]. Li *et al.* [48] enabled the re-rendered errors to be back-propagated to the light sources, refining them jointly with the normals simultaneously.

However, previous learning-based UPS methods rely on the explicit estimation of lighting directions to solve the GBR ambiguity. While end-to-end multi-view uncalibrated methods [49] leverage information from multiple viewpoints to achieve accurate 3D reconstruction, handling single-view

uncalibrated photometric stereo (UPS) in a single-stage framework remains an open challenge. As discussed in Section 1, estimating lighting directions and using them as input along with photometric stereo images may lead to training instability and complicated training steps, as well as the influence of discretized inaccurate lighting outputs. In contrast, we propose a new framework that uses a one-stage Fourier Embedding network to handle UPS, without the need for explicitly learning of lighting directions. Our method avoids the difficulty of explicitly learning exact lighting directions in a two-stage network and physically decomposes lighting and geometry information through the Fourier transform, thereby alleviating the GBR ambiguity.

3 Motivations

In this section, we provide more details to supplement the observation we highlighted in Section 1. Firstly, we briefly introduce the operation of the Discrete Fourier transform (DFT) [11], which represents an integral transform $\mathcal{F}$ that converts a spatial domain image O with resolution $H \times W$ into the frequency domain $\mathcal{O}$, as follows:

$$\begin{aligned}&\mathcal{F}(O(h,w)) = \mathcal{O}(u,v)\\&= \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} O(h,w) e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}\end{aligned} \tag{2}$$

where h, w and u, v represent the coordinates in the spatial domain and Fourier domain, respectively, and j is the imaginary unit. In the Fourier domain, $\mathcal{O}(u,v)$ is complex, containing real and imaginary components, as follows:

$$\mathcal{O}(u,v) = R(\mathcal{O}(u,v)) + jI(\mathcal{O}(u,v)), \tag{3}$$

where $R(\mathcal{O}(u,v))$ and $I(\mathcal{O}(u,v))$ represent the real and imaginary parts, respectively. Usually, the Fourier transform $\mathcal{O}(u,v)$ is represented by polar form, represented by the amplitude component $\mathcal{A}(\mathcal{O}(u,v))$ and the phase component $\mathcal{P}(\mathcal{O}(u,v))$, to provide an intuitive analysis, as follows:

$$\mathcal{A}(\mathcal{O}(u,v)) = \sqrt{R^2(\mathcal{O}(u,v)) + I^2(\mathcal{O}(u,v))}, \tag{4}$$

$$\mathcal{P}(\mathcal{O}(u,v)) = \arctan\left[\frac{I(\mathcal{O}(u,v))}{\mathcal{R}(\mathcal{O}(u,v))}\right]. \tag{5}$$

3.1 Lighting-geometry decomposition via DFT

Our main motivation arises from observing the relationship between photometric stereo images and the compositional images, which have their phase components swapped through DFT ($\mathcal{F}$) and IDFT ($\mathcal{F}^{-1}$). As shown in Fig. 1, when we swap the phases of two images with different illumination, the resulting compositional images almost preserve the original lighting cues, as O_{l_a} shares the same lighting with $\mathcal{F}^{-1}(\mathcal{A}(\mathcal{O}_{l_a}), \mathcal{P}(\mathcal{O}_{l_b}))$. Furthermore, when we swap the phase of the input images with different objects, the compositional images also interchange the objects (geometry), as O_{Buddha} shares the same geometry with $\mathcal{F}^{-1}(\mathcal{A}(\mathcal{O}_{Goblet}), \mathcal{P}(\mathcal{O}_{Buddha}))$. Therefore, we conclude that the lighting information and geometry information can be decomposed by the amplitude and phase in the Fourier domain, respectively, to a certain extent. In other words, we can obtain the lighting information and geometry information approximately from amplitude and phase in the Fourier domain, respectively.

In fact, our inference on photometric stereo images aligns with Fourier theory [50], [51], [52], [53], [54], [55], where the amplitude component reflects style information (*e.g.*, illumination characteristics), while the phase component represents semantic information (*e.g.*, the geometry itself). For example, in the Low-light Enhancement (LLE) task, the amplitude component is considered as the lightness of an image, while noises are revealed in the phase component [52], [54], [55]. Supported by this, DFT-based LLE methods enlarge the magnitude of its amplitude component and denoise the phase component via the neural networks, separately. Similarly, we follow the basic conclusion in previous works and extend it into the UPS task, in which illumination and geometry can be "decomposed" to a certain extent in the Fourier domain. Therefore, the implicit features of lighting can be extracted from the amplitude component, which enables the realization of the one-stage network for UPS via the Fourier transform.

3.2 Global information in DFT

Our second motivation stems from the definition of the Fourier transform as illustrated in Eq. (2). It can be observed that the Fourier transform $\mathcal{F}(O(h,w)) = \mathcal{O}(u,v)$ describes the image O in the frequency in terms of its amplitude and phase at each of its constituent frequencies. Every pixel in the spatial domain also contributes to a frequency component in the Fourier domain. That is to say, each component in the frequency domain contains some information from the entire spatial domain [11], [52], [54]. In the UPS task, $\mathcal{O}(u,v)$ contains global long-range spatial context cues such as cast shadows and inter-reflections under specific directions. These global cues are significant for UPS surface normal estimation. First, global features (*e.g.*, cast shadows) provide additional information on lighting directions. Second, the local shadows and specular highlights suffer from unreliable shading cues because of the under/overexposed values, which need the assistance of global information. Therefore, we will leverage the global properties of Fourier transform and combine them adaptively with local spatial information to recover further details.

4 Proposed Method

Based on the analysis in Section 3, we propose a deep Fourier Uncalibrated Photometric Stereo Network (FUPS-Net) with a one-stage pipeline, as shown in Fig. 2, leveraging the decomposition ability of the Fourier transform in photometric stereo images. In this section, we first introduce the structure and then delve into the details of each module.

As shown in Fig. 2, we design a two-branch structure consisting of the main branch and the auxiliary branch. The swapped compositional images in Fig. 1 are noisy and blurry, indicating that the decomposition in the Fourier domain is imperfect. Therefore, we also extract features from the spatial domain to assist in learning surface normals. The success of two-branch networks in various vision tasks [57], [58] is attributed to different focuses on its specific information processing procedure at different branches. By utilizing the distinct information from each branch in processing and appropriately combining them later, comprehensive information can be harnessed to significantly enhance surface

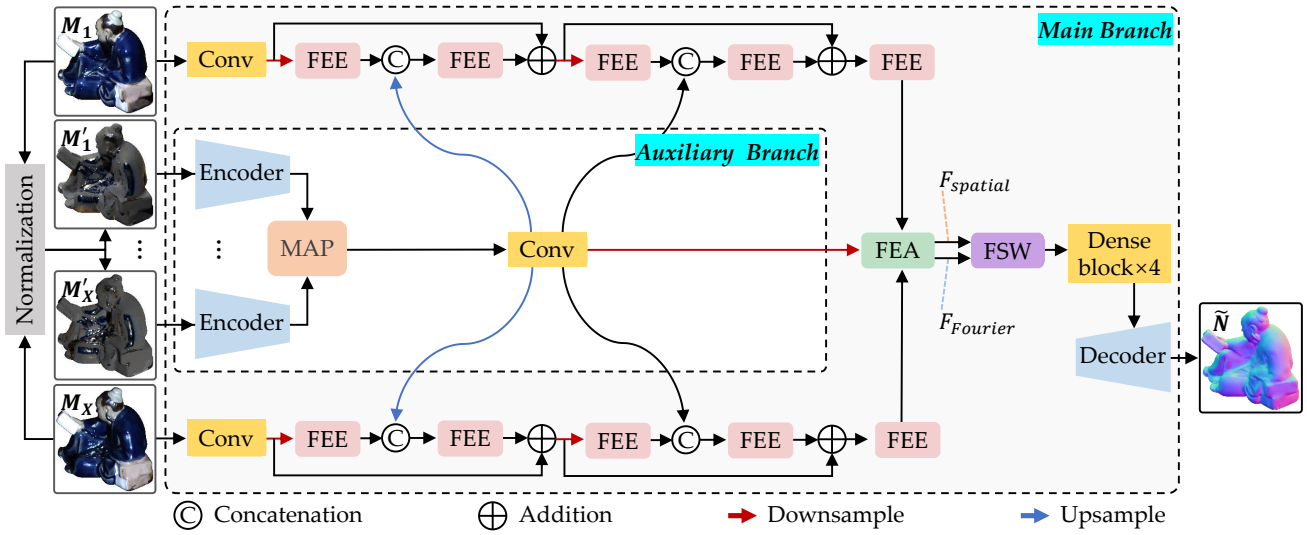


Fig. 2. Overview of FUPS-Net for surface normal estimation. The main branch takes the original images as input, which comprises five Fourier Embedding Extraction (FFE) blocks along with a Fourier Embedding Aggregation (FEA) block. Meanwhile, the auxiliary branch takes the normalized images as input and passes them to encoders, then aggregated by a Multi-head Attention Pooling (MAP) module [56]. Then, a frequency-spatial weighted (FSW) block is designed to adaptively combine the global Fourier frequency feature and local spatial feature, to further recover the details.

normal estimation performance. Furthermore, our FUPS-Net is a multi-input-single-output (MISO) network, because deep photometric stereo networks have to handle a variable number of input images.

The main branch of FUPS-Net takes original photometric stereo images $\boldsymbol{M_1}, \boldsymbol{M_2}, \ldots, \boldsymbol{M_X}$ as inputs, while the auxiliary branch uses normalized images $\boldsymbol{M_1'}, \boldsymbol{M_2'}, \ldots, \boldsymbol{M_X'}$ as inputs to assist the learning process. Motivated by the observation in Section 3, we propose the Fourier Embedding Extraction (FEE) block and the Fourier Embedding Aggregation (FEA) block for handling features in the Fourier domain. We detail these two key components in Sections 4.1 and 4.2. Specifically, the main branch comprises FEE blocks, which are organized in a residual manner [59], and incorporates two downsampling operations using bilinear interpolation. In addition, an FEA block is employed to manage a variable number of extracted frequency and spatial features. Subsequently, we propose using a frequency-spatial weighted (FSW) block to adaptively combine the global frequency feature and local spatial feature (see details in Section 4.3). Last, a 24-layer DenseNet module with four Dense blocks [60] is employed, followed by a Decoder structure to regress the estimated surface normals [20].

In the auxiliary branch, the normalization operation [9] is initially applied to alleviate the influence of spatially-varying BRDFs. This step is crucial as the CNN-based framework operates on patch-level inputs and is trained with a homogeneous BRDF. The encoder in the auxiliary branch has the same structure as the counterpart in PS-FCN [20]. However, the aggregation model, which merges a flexible number of features into one, differs from previous all-pixel-based photometric stereo networks [9], [20], [22], [35]. We introduce the Multi-head Attention Pooling (MAP) module [56], inspired by its applications in [25], [40]. This module enables us to reduce the number of elements in the set from an arbitrary dimension X to one by incorporating a learnable query Q, as opposed to solely retaining the maximum value as in [20]. Therefore, the MAP [56] serves as a global fusion method that considers all feature distributions for surface normal estimation.

Note that the FEE blocks in the main branch, along with the downsampling and 1×1 convolutional layer (Conv in Fig. 2) to adjust the spatial and channel dimensions, are concatenated with the features of the auxiliary branch at different scales, as illustrated in Fig. 2. This design facilitates the integration of global information in the spatial and channel dimensions. On the one hand, the output of the auxiliary branch maintains $1/4$ of the original resolution, while the concatenated features in the main features are $1/2$ and $1/4$ of the original resolution. Consequently, combining the output of the auxiliary branch with different receptive fields provides global information in the spatial domain. On the other hand, the output of the auxiliary branch aggregates features from all shading cues from different illumination directions, while the features in the main branch are extracted from a single photometric image. Therefore, combining the output of the auxiliary branch with the main branch features integrates information from both local and global cues, thereby enriching the channel domain. This approach enhances the network's capability to capture comprehensive information for surface normal estimation.

4.1 Fourier Embedding Extraction (FEE) Block

In Section 3, we found that geometry information and lighting information can be partially decomposed through the Fourier transform. Therefore, we propose using the FEE block for simultaneous feature extraction on amplitude and phase in the Fourier domain, inspired by the recent success of deep Fourier networks [52], [53], [54], [55], along with feature enhancement in spatial domain. As shown in Fig. 3 (a), the input features are split into the Fourier and spatial domains. Discrete Fourier Transform (DFT) is utilized to transform the input to the frequency domain, where the input is decomposed into its amplitude component ($\mathcal{A}$) and phase component ($\mathcal{P}$). These components then pass through two 3×3 convolutional layers with Leaky

ReLU activation, before being recombined using Inverse Discrete Fourier Transform (IDFT). In the spatial domain, we enhance the features using an efficient Half Instance Normalization (HIN) model [57], which is connected in parallel with a 3×3 convolutional layer, operating in a residual manner [59].

As discussed in Section 3, the processing of information in the Fourier domain allows for the capturing of global frequency representations, while convolutional layers primarily extract local representations in the spatial domain. Therefore, our FEE block utilizes an interactive method to combine these two representations. Specifically, we interact with the output features from the first Fourier domain operator F_{f1} and the first spatial domain operator F_{s1} as follows:

$$F_{f2_{in}} = F_{f1} + \mathcal{W}_s(F_{s1}), \tag{6}$$

$$F_{s2_{in}} = F_{s1} + \mathcal{W}_f(F_{f1}), \tag{7}$$

where $\mathcal{W}_s(\cdot)$ and $\mathcal{W}_f(\cdot)$ represent a 3×3 convolutional layer, and $F_{f2_{in}}$ and $F_{s2_{in}}$ represents the interacted features for the second Fourier domain operator and spatial domain operator, respectively. The subsequent operations follow the same formulation as the first ones. Finally, we concatenate them followed by 1×1 convolutional layer to adjust the channel dimensions. To further enhance feature representation, we employ skip connections (element-wise addition) to combine the input feature with the output, creating a residual structure [59].

4.2 Fourier Embedding Aggregation (FEA) Block

As previously mentioned, FUPS-Net is a MISO network because photometric stereo needs to handle a variable number of input images. Managing this variability necessitates an additional fusion model to consolidate variable amount features into a representation with a fixed number of channels. This requirement arises because CNN-based networks lack intrinsic capabilities to manage variable number of inputs during both training and testing phases [9]. To address this limitation, we further propose the Fourier Embedding Aggregation (FEA) block to output aggregated features with a fixed number of channels for backpropagation.

As shown in Fig. 2 and Fig. 3 (b), each main branch feature is concatenated with the global feature from the auxiliary branch before processing. The aim of this design is to enhance the fusion of global information across both spatial and channel dimensions and to mitigate information loss during aggregation operations. Each input is decomposed into $\mathcal{A}_i$ and $\mathcal{P}_i$, $i \in \{1, 2, \cdots, X\}$, suding DFT. Subsequently, these components are subjected to two 3×3 convolutional layers with Leaky ReLU activation (similar to FEE block). MAP [56] is then adopted for $\mathcal{A}_1$ to $\mathcal{A}_X$, and $\mathcal{P}_1$ to $\mathcal{P}_X$. Compared to the previous max pooling strategy [20], it serves as a global fusion method that considers all feature distributions. The aggregated amplitude and phase features from MAP are further aggregated by using IDFT. In the FEA block, we also incorporate information from the spatial domain to complement the information in the Fourier domain, which shares the same structure as that in the FEE block and is also aggregated using MAP. Consequently, we eventually obtain two aggregated features from the FEA block: the spatial feature $F_{spatial}$ and the frequency feature $F_{Fourier}$.

4.3 Frequency-spatial Weighted (FSW) block

The FEA block outputs two aggregated features: one from the spatial domain, denoted as $F_{spatial}$, and the other from the Fourier domain, denoted as $F_{Fourier}$. As depicted in Eq. (2), each single feature in the Fourier domain encompasses information from the entire spatial domain. Consequently, $F_{Fourier}$ captures global information, including long-range context cues such as shadows and inter-reflections. Given that photometric stereo images may contain shadows and inter-reflections that influence local shading cues, capturing long-range context becomes crucial for accurate feature extraction. To fully leverage both the global properties of frequency information and the local features of spatial information, we introduce the Frequency-Spatial Weighted (FSW) block. This block adaptively combines $F_{spatial}$ and $F_{Fourier}$ to optimize feature integration.

In the FSW block, we first design a Normalized High-frequency (NHF) operation to calculate the NHF map, which serves as the weight to merge the global frequency branch and the local spatial branch. Specifically, the NHF map Ω is computed as follows:

$$\hat{\boldsymbol{M}}_{\boldsymbol{i}}' = \mathrm{blur}(\boldsymbol{M}_{\boldsymbol{i}}'), \tag{8}$$

$$\boldsymbol{D}_{\boldsymbol{i}} = \mathrm{abs}(\boldsymbol{M}_{\boldsymbol{i}}' - \hat{\boldsymbol{M}}_{\boldsymbol{i}}'), \tag{9}$$

$$\boldsymbol{\Omega}_{\boldsymbol{i}} = \hat{\boldsymbol{M}}_{\boldsymbol{i}}'./\boldsymbol{D}_{\boldsymbol{i}}, \tag{10}$$

$$\boldsymbol{\Omega} = \mathrm{average}(\boldsymbol{\Omega}_{\mathbf{1}}, \boldsymbol{\Omega}_{\mathbf{2}}, \ldots, \boldsymbol{\Omega}_{\boldsymbol{X}}), \tag{11}$$

where $\boldsymbol{M}_{\boldsymbol{i}}'$ represents the i_{th} normalized photometric stereo image [9], $\mathrm{blur}(\cdot)$ denotes Gaussian blur with an empirically set kernel size of 9×9, $\mathrm{abs}(\cdot)$ computes the absolute value, $(\cdot)./(\cdot)$ performs elementary division, and $\mathrm{average}(\cdot)$ stands for average pooling aggregation. Consequently, $\hat{\boldsymbol{M}}_{\boldsymbol{i}}'$ denotes the image blurred via the Gaussian blur kernel, exhibiting increased blur effect in high-frequency regions, such as specular highlights, shadows, and crinkle structures. As a result, the absolute difference $\boldsymbol{D}_{\boldsymbol{i}}$ in these regions has larger values, while smaller values in its counterparts $\boldsymbol{\Omega}_{\boldsymbol{i}}$. Eqs. (8), (9), and (10) constitute the NHF operation, as depicted by the purple box in Fig. 4. To handle the flexible number of input images and integrate all information, we further utilize average pooling to aggregate an arbitrary number of $\boldsymbol{\Omega}_{\mathbf{1}}, \boldsymbol{\Omega}_{\mathbf{2}}, \ldots, \boldsymbol{\Omega}_{\boldsymbol{X}}$ into a single NHF map, denoted as $\boldsymbol{\Omega}$. Lower values in $\boldsymbol{\Omega}$ correspond to surface positions with more long-range-related regions, requiring additional global information for reference. Conversely, higher values in $\boldsymbol{\Omega}$ represent flat or clear shading cue regions with fewer global effects. Therefore, we argue that regions characterized by lower values in the NHF map tend to be processed within the Fourier feature $F_{Fourier}$, while regions with higher values in the NHF map are inclined to be processed within the local spatial feature $F_{spatial}$. This is further demonstrated in the object "Reading" in Fig. 4, where the dark regions in $\boldsymbol{\Omega}$ represent the areas of cast shadow and specular highlight.

Consequently, as illustrated in Fig. 4, we propose to integrate the global Fourier feature and the local spatial feature based on the NHF map $\boldsymbol{\Omega}$ (normalized to [0, 1]) and

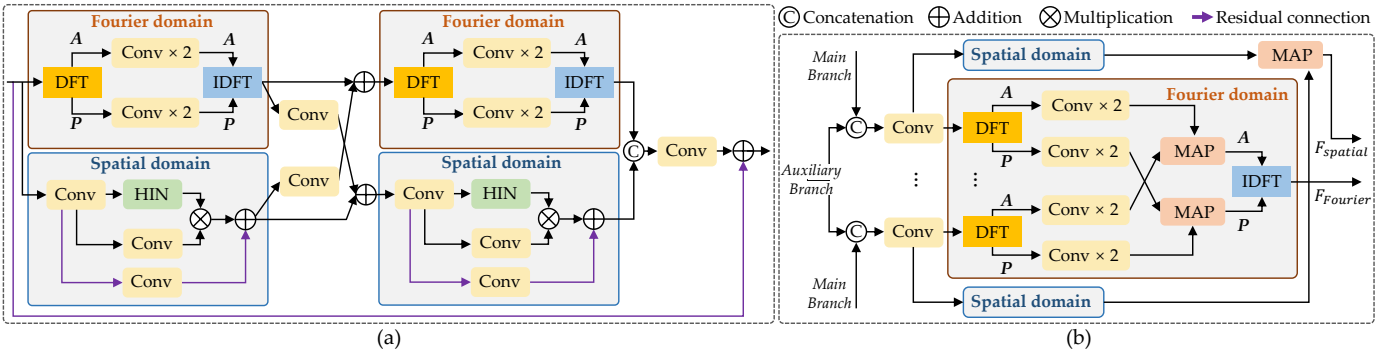


Fig. 3. Structure of the proposed (a) Fourier Embedding Extraction (FEE) block, and (b) Fourier Embedding Aggregation (FEA) block. $\mathcal{A}$ and $\mathcal{P}$ represent the amplitude component and phase component, respectively. HIN stands for the Half Instance Normalization model [57].

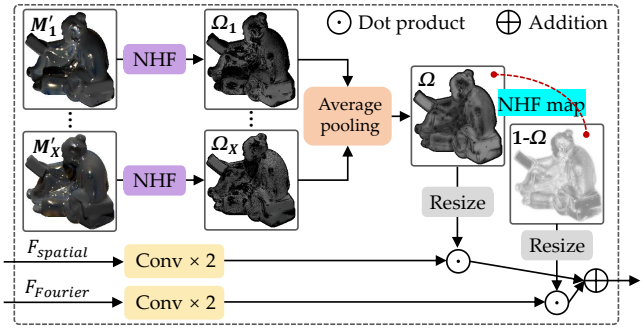


Fig. 4. Structure of the proposed Frequency-Spatial Weighted (FSW) block. NHF stands for the Normalized High-frequency.

its counterpart $1-\mathbf{\Omega}$. Before integration, the features $F_{spatial}$ and $F_{Fourier}$ undergo two 3×3 convolutional layers with Leaky ReLU activation, and the maps $\mathbf{\Omega}$ and $1-\mathbf{\Omega}$ are down-sampled to match the size of the features $F_{spatial}$ and $F_{Fourier}$. The integration can be expressed as follows:

$$F_{int} = \mathbf{\Omega} \odot F_{spatial} + (1-\mathbf{\Omega}) \odot F_{fourier}, \quad (12)$$

where F_{int} is the output feature. Finally, as shown in Fig. 2, we employ a 24-layer dense connected module comprising four Dense blocks [60], followed by a decoder to regress the estimated surface normals with an L2-normalization layer [9].

4.4 Learning procedures

During training, we optimize the proposed FUPS-Net by minimizing the following loss function $\mathcal{L}$, as follows:

$$\mathcal{L} = \|1 - \tilde{\boldsymbol{N}} \odot \boldsymbol{N}\|_1 + 0.1 \times \|\mathrm{VGG}(\tilde{\boldsymbol{N}}) - \mathrm{VGG}(\boldsymbol{N})\|_2, \quad (13)$$

where $\boldsymbol{N}$ represents the ground truth and $\tilde{\boldsymbol{N}}$ stands for the estimated surface normals. The first term denotes the commonly used cosine similarity loss, with the symbol $\odot$ representing the dot product operation. If the estimated surface normals $\tilde{\boldsymbol{N}}$ exhibit a similar orientation to the ground-truth $\boldsymbol{N}$, $\tilde{\boldsymbol{N}} \odot \boldsymbol{N}$ will approach 1, causing the first term to tend towards 0. In the second term, we incorporate a perceptual loss to enhance high-frequency details [61], with the weight factor empirically set to 0.1. The perceptual loss is computed using the pre-trained VGG-19 network, which is supervised at four scales.

FUPS-Net is implemented using PyTorch. We employ the Adam optimizer with default settings ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) on an RTX 4090 GPU with 24GB memory. The initial learning rate is set to 0.002 and halved every 5 epochs. We train FUPS-Net using a batch size of 32 for 40 epochs. During training, we utilize 32 input images, each sized at 64 $\times$ 64. It takes approximately 20 hours to train our FUPS-Net and 9 seconds to test one object in the DiLiGenT dataset [27], with 96 input images. Note that the number of input images and their size can be flexibly adjusted during testing. Our network is trained on publicly available synthetic Blobby and Sculpture shape datasets [62], utilizing rendered photometric stereo images provided by [9]. The dataset consists of 85,212 samples, with each sample comprising 64 images captured from various lighting directions sampled from the upper hemisphere. Among these images, 99%, or a total of 84,362 samples, are utilized for training FUPS-Net, while the remaining 852 samples are employed for validation purposes.

5 Experimental Results

To verify the quantitative accuracy of the estimated surface normals, we use the mean angular error (MAE) in degrees, calculated as follows:

$$\mathrm{MAE} = \frac{1}{U} \sum_{p}^{U} \cos^{-1}\left(\tilde{\boldsymbol{n}}^p \cdot \boldsymbol{n}^p\right), \quad (14)$$

where U is the total number of pixels in the surface area, and $\tilde{\boldsymbol{n}}_{\boldsymbol{p}}$ and $\boldsymbol{n}_{\boldsymbol{p}}$ are the surface-normal vector at pixel p of the ground-truth $\tilde{\boldsymbol{N}}$ and the estimated surface normals $\boldsymbol{N}$, respectively.

5.1 Ablation Studies

We conduct ablation studies to analyze the effectiveness of the main components of our design. Table 1 presents the quantitative comparison of the ablated models on the validation set. We report the average MAE across 852 samples, each with 64 input images. Our complete model is denoted as #0. First, we ablate the FEE block, including removing the Fourier domain representation (#1), removing the spatial domain representation (#2), and utilizing the Fourier and spatial domain representations once without interactive combination (#3). We then evaluate the FEA block by removing the Fourier domain representation(#4), removing the spatial domain representation (#5), excluding the concatenation of the global feature from the auxiliary

TABLE 1
Quantitative comparison of ablation studies on our FUPS-Net, in terms of average MAE in degrees, on the validation set, where FD stands for Fourier domain, SD for Spatial domain, AB for Auxiliary branch, AP for Average pooling, MP for Max pooling, $\mathcal{A}$ for Amplitude component, and $\mathcal{P}$ for Phase component.

#	FEE block		FEA Block			FSW Block		MAE (°) ↓
	FD	SD	FD	SD	AB	AP	MP	
0	✓	✓	✓	✓	✓	✓		5.36
1		✓	✓	✓	✓	✓		5.87
2	✓		✓	✓	✓	✓		5.52
3	1 ×	1 ×	✓	✓	✓	✓		5.59
4	✓	✓		✓	✓	✓		6.81
5	✓	✓	✓		✓	✓		5.74
6	✓	✓	✓	✓		✓		5.77
7	✓	✓		✓		✓		6.89
8	✓	✓	✓			✓		5.95
9	✓	✓	✓	✓	✓		✓	5.50
10	✓	✓	✓	✓	✓			5.73
11	$\mathcal{A}$	✓	$\mathcal{A}$	✓	✓			6.39
12	$\mathcal{P}$	✓	$\mathcal{P}$	✓	✓			7.15
13	$\mathcal{A}$		$\mathcal{A}$					47.73
14	$\mathcal{P}$		$\mathcal{P}$					15.41
15		✓		✓				7.64

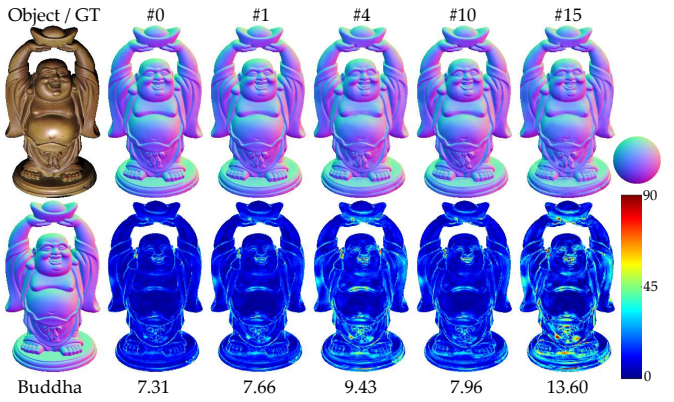


Fig. 5. Visualization of the key ablation studies based on the object Buddha of the DiLiGenT benchmark [27]. The contrast of the images has been adjusted to improve visualization.

branch (#6), removing the Fourier domain representation without concatenating the global feature (#7), and removing the spatial domain representation without concatenating the global feature (#8). Furthermore, we test the FSW block using max pooling aggregation (#9) and without the FSW block (#10). In addition, we experiment with features in the Fourier domain based on the single amplitude component (#11, #13) and phase component (#12, #14), with or without the spatial domain features. Finally, we measure the performance when all modules related to the Fourier domain and auxiliary branch are discarded (#15). We also visualize some key ablation experiments on the object Buddha from the DiLiGenT benchmark dataset [27], with all 96 input images, as shown in Fig. 5.

As shown in Table 1, we can see that all the key designs contribute to the optimal performance achieved by the full model. Overall, the absence of the Fourier domain (#1 and #4) results in a drop in MAE (0.51° and 1.45°), highlighting the crucial role of decomposing amplitude and phase in the Fourier domain to enhance surface normal estimation when dealing with unknown lighting directions. Compared with #1, the aggregation of Fourier domain features (#4) is much more significant for the performance of UPS. Furthermore, the results of #2 and #5 demonstrate that the spatial domain also contributes to boosting the accuracy of surface normal estimation, showing the complementary information it contains. In fact, without the Fourier domain in both the FEE and FEA blocks (#15), our method regresses to an original one-stage UPS network similar to UPS-FCN [20], which exhibits unsatisfactory performance. This demonstrates that Fourier decomposition can implicitly extract lighting information and geometry information from the amplitude and phase spectra, respectively. Consequently, the GBR ambiguity can be resolved in a one-stage network without explicit input for lighting directions.

Specifically, in the FEE block, we further test the results with only one Fourier domain and spatial domain processing (#3), which means the interactive combinations (Eqs. (6) and (7)) are discarded. The increased error (MAE) observed in this setup illustrates that the combined global frequency representation and local spatial representation also contribute significantly to feature extraction.

Furthermore, in the FEA block, we evaluate the effect of the aggregated global feature from the auxiliary branch. The global feature serves two purposes: (1) it offers global shading cues from all input illumination directions, and (2) it provides normalized shading cues unaffected by spatially-varying BRDFs. #6 illustrates how the concatenation of the global feature benefits the reconstruction results. #7 and #8 further show the effect of the global feature for spatial domain and Fourier domain aggregations, respectively. Compared with #4 and #5, we can see that the global feature is more effective with Fourier domain aggregation, as evidenced by the larger degradation between #5 and #8 (0.21°) than between #4 and #7 (0.08°). These results indicate that the global feature fused in the auxiliary branch is also a spatial feature that shares similar effects with the spatial domain aggregation in the FEA block. Ablations #4 - #8 reflect that learning the Fourier domain feature plays an important role in the decomposition of unknown illuminations and surface geometry, rather than relying solely on spatial features as previously used.

In addition, ablations with #9 and #10 test the effect of the FSW block. In #9, the performance of FUPS-Net decreases with the use of max pooling operation to fuse $\mathbf{\Omega_1}, \mathbf{\Omega_2}, \ldots, \mathbf{\Omega_n}$ in Eq. (11). This might be explained by the fact that shadows and inter-reflections exhibit imbalanced distributions in features from different illuminations, and max pooling disregards these imbalances. This occurs because max pooling only retains the maximum value at the same position while average pooling records the global representation to some extent. In #10, we cancel both average pooling and max pooling, resulting in the aggregated $F_{spatial}$ and $F_{Fourier}$ being simply concatenated without the FSW block. It can be observed that our FUPS-Net performed better when utilizing the FSW block to adaptively combine the global Fourier domain infomration and local spatial information. It is interesting to note that for complex objects with more cast shadows and inter-reflections, such as the Buddha in Fig. 5, discarding the FSW block may lead to

a larger decrease in performance. For example, #1 performs worse than #10 in the validation set, but it is opposite for the object Buddha with a very complex structure. It further suggests that the FSW block can facilitate the fusion of long-range global context and local information.

Finally, experiments #11 - #14 illustrate the effect of a single component in the Fourier domain. Specifically, #11 represents the ablation discarding the phase component, with the extracted features from the spatial domain, and #12 does the same for the amplitude component. Experiments #13 and #14 further test the corresponding ablations without fusing the spatial domain features. We processed each component individually in the Fourier domain and transformed the one component feature to the spatial domain representation via IDFT [53]. Note that the FSW block is not utilized because incomplete Fourier domain features make the frequency-spatial fusion meaningless. It is evident that only the amplitude feature involved (#11) achieves better performance compared to the phase component (#12) when spatial domain features are fused. This suggests that the amplitude component contains implicit lighting information, which guides the resolution of GBR ambiguity in the spatial features. However, discarding all spatial features will cause significant degradation in performance if only a single Fourier component is maintained. In fact, discarding the phase component alone (#13) can even lead to the non-convergence of our method. These results highlight that the single amplitude component lacks crucial geometry information, and fails to constrain the learning of surface normals without the phase component or spatial features combined.

5.2 Evaluation on Benchmark Datasets

We first evaluate our FUPS-Net and compare it with previous calibrated and uncalibrated methods on the widely used photometric stereo dataset, namely the DiLiGenT benchmark [27]. DiLiGenT contains 10 objects and each object has 96 images captured under different lighting conditions. The quantitative results for surface normal estimation are tabulated in Table 2. Additionally, Fig. 6 provides visual representations of the reconstruction results and error map comparisons for the Reading and Harvest objects.

As shown in Table 2, our FUPS-Net achieves the best results on the UPS task and outperforms most CPS methods, concurrent with the SCPS-NIR method [48], in terms of the average MAE of the ten objects in DiLiGenT. It can be seen that our FUPS-Net can achieve the best or sub-optimal results on all objects in the DiLiGenT dataset. FUPS-Net performs well on most objects with complex structures and strong non-Lambertian surface reflectances, such as Buddha, Cow, Harvest, and Reading. However, it is noteworthy that for very simple structured objects such as Ball and Bear, SCPS-NIR [48] demonstrates more reasonable results. This could be attributed to two reasons: (1) these objects are easy to acquire supervision by neural inverse rendering with jointly optimized object shape and lighting information, and (2) the simple structures may lead to inefficient feature extraction from the decomposed amplitude and phase components in the Fourier domain. Note that the first 20 images are photometrically inconsistent in the belly region of the object Bear [10]. When discarding the first 20 images, the results of our FUPS-Net achieves 5.08° on Bear and 7.03° on average MAE. As shown in Fig. 6, the proposed FUPS-Net outperforms previous methods in the regions with specular inter-reflections (cloth of the object Reading), cast shadows (pocket of the object Harvest), and crinkle surfaces (cloth of the object Harvest), further demonstrating the effectiveness of the introduced Fourier domain decomposition and FSW block.

On the other hand, it is worth emphasizing that the proposed FUPS-Net is the first one-stage end-to-end UPS network with reasonable surface-normal estimation performance. Compared with the previous two-stage methods that take the first-stage estimated lighting directions and intensities as the input of the second-stage surface normal network, our approach eliminates the need for multi-stage training and avoids the influence of inaccurately classified lighting directions. For instance, the two-stage method SDPS-Net [21] almost costs double the training time to train LCNet and NENet separately, compared to our one-stage FUPS-Net. Furthermore, the two-stage methods introduce instability in surface learning through the conversion of lighting direction estimation from continuous regression to discrete classification, while our implicit one-stage network avoids error accumulation in the previous two-stage UPS network. Compared with neural inverse rendering methods, the computational cost of FUPS-Net scales almost linearly with the number of input images, similar to CPS networks. For instance, our FUPS-Net only takes no more than 10 seconds to test one object in the DiLiGenT dataset [27], while it costs approximately 15 minutes for the concurrent SOTA method SCPS-NIR [48], and 50 minutes for the two-stage neural inverse rendering method SK21 [23]. This characteristic makes FUPS-Net significantly more efficient than the neural inverse rendering-based methods and two-stage methods in terms of computational resources.

Generally, the universal photometric stereo environment can also be seen as a kind of UPS task, which is not limited to the assumption of any specific lighting conditions, *e.g.*, directional lighting. We also compare our method to the two recent Universal photometric stereo methods UniPS [25] and SDM-UniPS [26]. The compared results are summarized in Table 3. Although our method achieves sub-optimal results on the average MAE of ten objects from the DiLiGenT dataset [27], performing slightly worse than SDM-UniPS, it demonstrates clear advantages in terms of model complexity and computational efficiency. Specifically, SDM-UniPS employs a network with 125.73 million parameters and requires 72 hours of training on an A100 GPU, while FUPS-Net uses a much lighter network with only 11.39 million parameters and completes training in just 20 hours on an RTX 4090 GPU. Furthermore, FUPS-Net is trained using a dataset with 1.38 billion total pixels, which is considerably smaller than the 9.15 billion pixels used by SDM-UniPS. Notably, SDM-UniPS cannot run tests on the DiLiGenT dataset using all 96 input images on an RTX 4090 GPU with 24GB memory, whereas FUPS-Net can be executed on less restrictive hardware. These results highlight that FUPS-Net offers significant benefits in terms of efficiency and resource requirements, making it more practical for scenarios with limited computational resources, while maintaining com-

TABLE 2
Performance of different methods on the DiLiGenT benchmark [27] with 96 images, in terms of MAE in degrees. UPS-1s stands for the one-stage methods without explicitly learning lighting information, and UPS-exp stands for the previous methods (either two-stage or neural inverse rendering) that need to explicitly learn the lighting directions. For the UPS task, we use **bold font** and underline to highlight the best and second-best results, respectively.

Method	Task	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg. ↓
IRPS [63]	CPS	1.47	5.79	10.36	5.44	6.32	11.47	22.59	6.09	7.76	11.03	8.83
PS-FCN [20]	CPS	2.82	7.55	7.91	6.16	7.33	8.60	15.85	7.13	7.25	13.33	8.39
GPS-Net [36]	CPS	2.92	5.07	7.77	5.42	6.14	9.00	15.14	6.04	7.01	13.58	7.81
CNN-PS [10]	CPS	2.12	8.30	8.07	4.38	7.92	7.42	14.08	5.37	6.38	12.12	7.62
PS-FCN$^{\text{Norm.}}$ [9]	CPS	2.67	7.72	7.53	4.76	6.72	7.84	12.39	6.17	7.15	10.92	7.39
NormAttention-PSN [35]	CPS	2.93	5.48	7.12	4.65	5.99	7.49	12.28	5.96	6.42	9.93	6.83
LL22 [64]	CPS	2.43	3.64	8.04	4.86	4.72	6.68	14.90	5.99	4.97	8.75	6.50
PX-Net [65]	CPS	2.03	4.13	7.61	4.39	4.69	6.90	13.10	5.08	5.10	10.26	6.33
UPS-FCN [20]	UPS-1s	6.62	11.23	15.87	14.68	11.91	20.72	27.79	13.98	14.19	23.26	16.02
DeepPS2 [66] (input = 2)	UPS-exp	6.28	9.67	14.51	9.87	11.08	14.22	26.06	10.73	12.09	19.94	13.44
KS21 [23]	UPS-exp	3.78	5.96	13.14	7.91	10.85	11.94	25.49	8.75	10.17	18.22	11.62
SDPS-Net [21]	UPS-exp	2.77	6.89	8.97	8.06	8.48	11.91	17.43	8.14	7.50	14.90	9.51
SK22 [22]	UPS-exp	3.46	5.48	10.00	8.94	6.04	9.78	17.97	7.76	7.10	15.02	9.15
UPS-GCNet [24]	UPS-exp	2.50	5.60	8.60	7.80	8.48	9.60	16.20	7.20	7.10	14.90	8.70
LERPS [67]	UPS-exp	2.41	6.93	8.84	7.43	6.36	8.78	**11.57**	8.32	7.01	11.51	7.92
SCPS-NIR [48]	UPS-exp	**1.24**	**3.82**	9.28	4.72	**5.53**	**7.12**	14.96	6.73	**6.50**	**10.54**	7.05
FUPS-Net (Ours)	UPS-1s	2.30	5.27	**7.31**	**4.71**	5.74	7.54	13.75	**6.03**	6.58	11.29	7.05

When discarding the first 20 images of the object Bear, our FUPS-Net achieves 5.08° on Bear and 7.03° on average MAE.

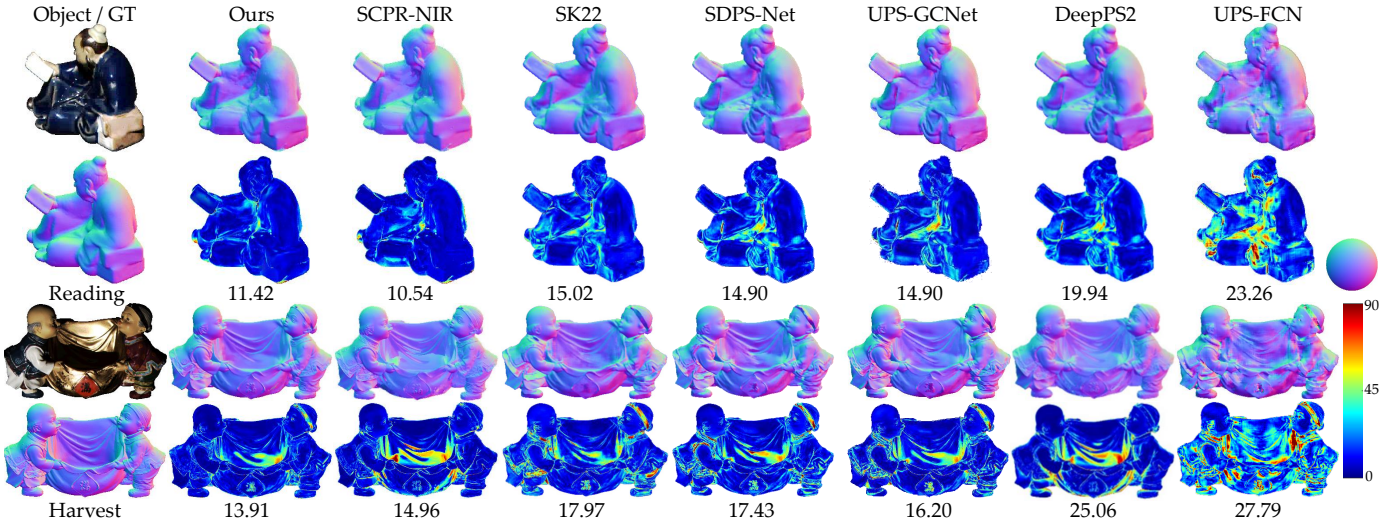


Fig. 6. Quantitative results on the DiLiGenT dataset [27] with 96 input images. In each sample, the first row displays the estimated normal maps, while the second row depicts the error maps obtained from various methods. The values indicate the MAE in degrees. The contrast of the images has been adjusted to improve visualization.

TABLE 3
Comparisons with two universal photometric stereo methods, on the DiLiGenT benchmark [27] with average MAE in degrees, training dataset (number of total pixels), network parameters, and running time.

Method	Avg. MAE	Dataset	Parameters	Time
UniPS [25]	14.70	2.65B	69.40M	48h (RTX 8000)
SDM-UniPS [26]	5.80	9.15B	125.73M	72h (A100)
FUPS-Net (Ours)	7.05	1.38B	11.39M	20h (RTX 4090)

petitive performance.

To conduct a comprehensive analysis of the generalization capability of our FUPS-Net across various objects and materials, we further use the challenging DiLiGenT10^2 dataset [28]. DiLiGenT10^2 contains 100 objects of 10 shapes multiplied by 10 materials and each object has 100 images under different conditions. These datasets pose significant challenges due to their inclusion of strongly non-Lambertian surface materials and complex structures. The results, as obtained from the online evaluation website, are presented in Fig. 7. We further present the visualized results of all ten shapes based on the worst-performing material ACRYLIC in Fig. 8. Since ground truth surface normals are not provided, we utilize the method proposed by [68] to integrate the estimated surface normals into the 3D reconstruction to intuitively illustrate the performances. For more results, please refer to https://github.com/Kelvin-Ju/FUPS-Net.

As shown in Fig. 7, we can see that FUPS-Net obviously outperforms the similar one-stage UPS-FCN [20] and achieves comparable accuracy to calibrated PS-FCN$^{\text{Norm.}}$ [9] on most samples. Fig. 8 further illustrates the improvement of the existing one-stage UPS method. The previous one-stage UPS-FCN [20] fails to adequately reconstruct the shape

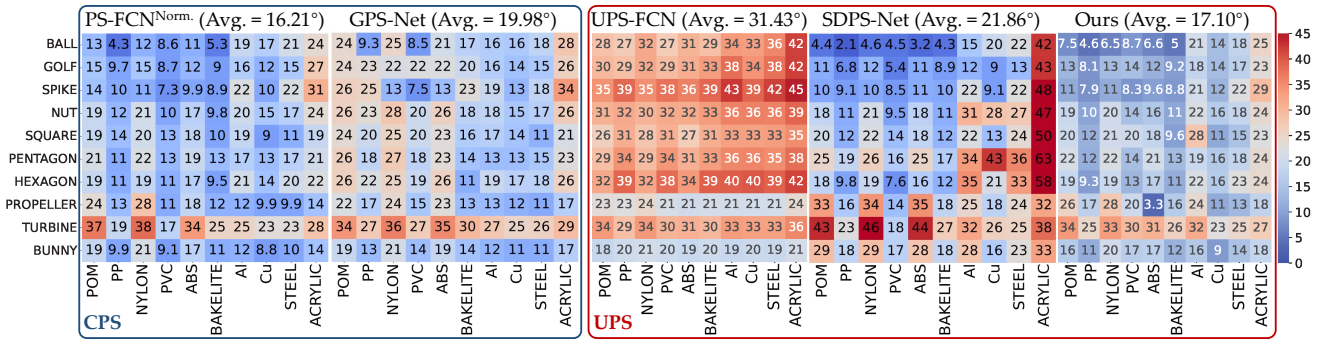


Fig. 7. The shape-material error matrix used to compare our FUPS-Net with recent calibrated and uncalibrated methods. The number in each element of the matrix represents the MAE in degrees according to the shape and material index.

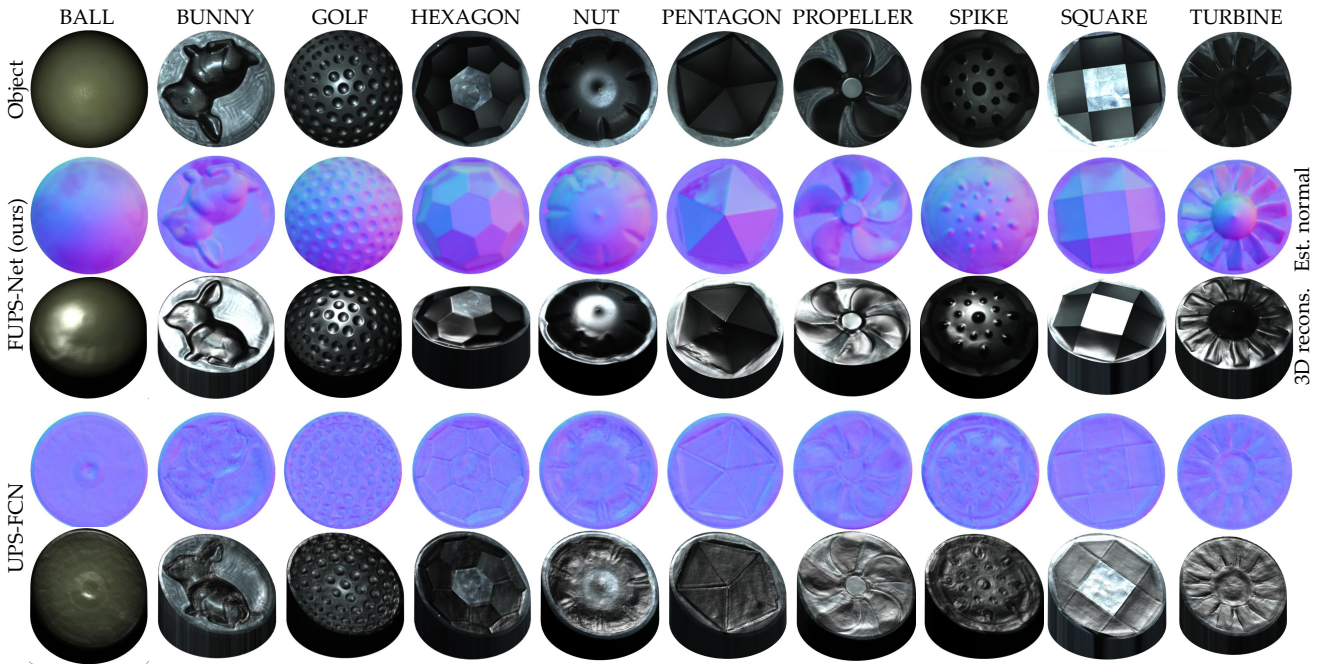


Fig. 8. Visualization results of our FUPS-Net and UPS-FCN [20] of all ten shapes in the DiLiGenT10^2 dataset [28] based on the worst-performing material ACRYLIC. The 3D reconstruction (3D recons.) results of estimated surface normal maps (Est. normal) are illustrated using [68].

of the objects, whereas our method successfully estimates detailed surface structures and provides reasonable 3D reconstruction results. These demonstrate that the implicitly extracted features by Fourier decomposition in our method can effectively address the geometry-lighting ambiguity, known as the GBR ambiguity [16] in photometric stereo under uncalibrated lighting directions.

5.3 Evaluation on Other Datasets

In this section, we first evaluate our method using the synthetic object Armadillo from the Stanford 3D dataset [29]. The Armadillo shape exhibits intricate surface structures and is rendered using MERL BRDFs [69] with 100 different materials, similar to the materials in our training dataset. Each material is illuminated by 100 random lighting directions from the upper hemisphere. Fig. 9 displays the MAE of predicted normal maps for the Armadillo object across the 100 materials, sorted by their MAE values.

As shown in Fig. 9, the proposed FUPS-Net demonstrates promising results across 100 different materials, achieving an average MAE of 9.12°. It is evident that our method robustly handles most surface materials, with only three materials exhibiting errors exceeding 12 degrees, which have significantly strong non-Lambertian properties (such as example J). To further analyze the reason for this, we visualize the Fourier transforms of these samples, denoted as A (pure-rubber), B (yellow-phenolic), E (gold-paint), I (nickel), and J (chrome), in Fig. 10.

As shown in Fig. 10, the amplitude component of materials showcases different representations due to the coupling of lighting information with materials. Non-Lambertian materials (I and J) exhibit obvious specular highlights, while Lambertian-related materials (A and B) present continuous shading cues. For these five examples with the same surface normal (3D structure), we can see the amplitude components are changed, while the phase components show almost the same. This observation aligns with the findings in Fig. 1, indicating that lighting and geometry information are decomposed into the amplitude and phase in the Fourier domain to some extent, respectively. Moreover, materials I and J display sparse and global noise across the entire spectrum, while materials A and B exhibit clearer and more concentrated features in their amplitude spectra. It may suggest that an amplitude feature with sparse and globally

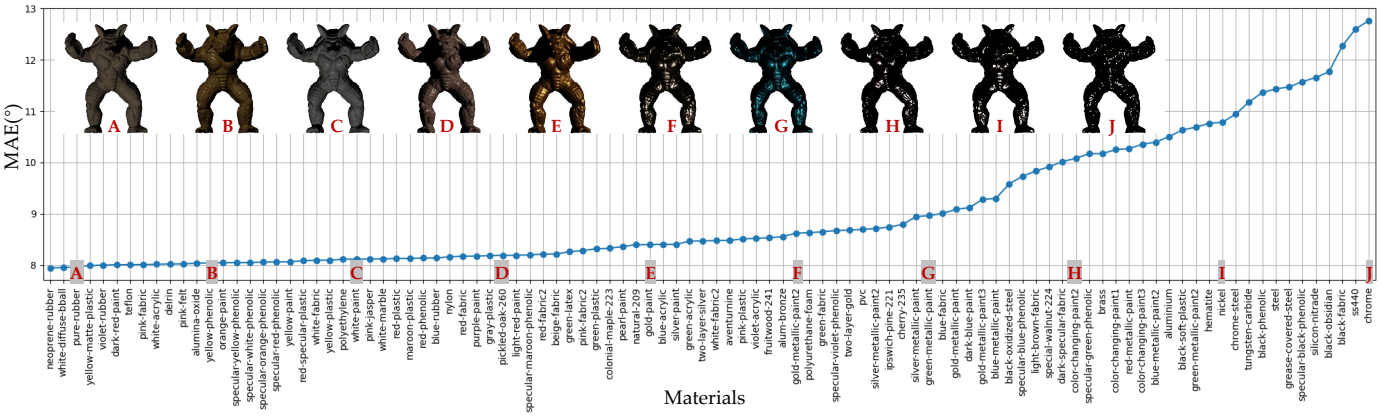


Fig. 9. MAE of the predicted surface normals of our FUPS-Net for the Armadillo object across 100 materials in the MERL BRDF dataset [69]. Some input examples are shown at the top.

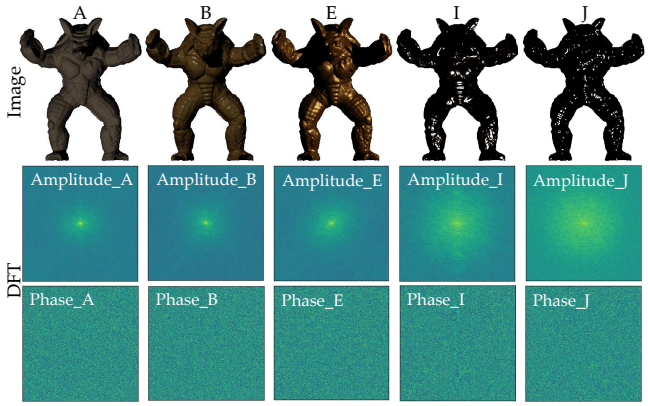


Fig. 10. Visualization of five material examples corresponding to Fig. 9. Materials A, B, and E exhibit good reconstruction results, while I and J report relatively poorer performance. We employ DFT to generate their amplitude and phase components, respectively.

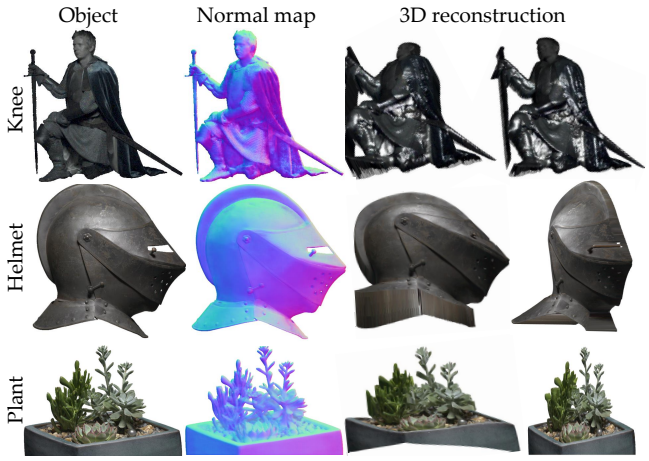


Fig. 11. Evaluation on the Light Stage Data Gallery, with only 32 input images. The estimated surface normals are shown qualitatively. The 3D reconstruction results of our estimated surface normal maps are also illustrated using [68]. The contrast of the images is adjusted for easier visualization.

distributed noise could present challenges in extracting implicit lighting information and cause worse performance on surface-normal estimation.

We further evaluate our method using the more intricate Light Stage Data Gallery dataset [70], which incorporates general non-Lambertian materials, complex structures, and lower-quality images. As ground truth data is unavailable for this dataset, we present qualitative results for the objects Knee, Helmet, and Plant using our FUPS-Net in Fig. 11. These results encompass surface normals and 3D reconstruction results obtained via [68], utilizing 32 randomly selected input images from a pool of 253 images.

As shown in Fig. 11, our FUPS-Net achieves detailed surface reconstruction on these objects, such as the screw of the object Helmet. We also note that the number of input images for each object in the Light Stage Data Gallery dataset [70] is much less compared to the above benchmark datasets [27], [28], demonstrating the robustness of our method when dealing with sparse input images. However, the distortion observed in the 3D reconstruction of object Knee in Fig. 11 can be attributed to the unique challenges posed by the object itself. The object Knee exhibits significant self-occlusions and non-Lambertian surface properties, which make accurate reconstruction inherently difficult. Furthermore, the noticeably distorted region, the head of Knee, is located far from the object's center, effectively placing it closer to the light source relative to the object's central regions. This relative proximity amplifies the influence of near-field lighting effects, making it more susceptible to issues commonly associated with close-range photometric stereo problems, such as non-uniform light distribution and perspective distortions.

6 Conclusion

In this paper, we propose a Fourier transform-based one-stage uncalibrated photometric stereo (UPS) method, namely FUPS-Net. Our approach is motivated by the observation that the lighting and geometry information can be "decomposed" in the Fourier domain to some extent. Leveraging this insight, FUPS-Net incorporates the proposed Fourier Embedding Extraction (FEE) and Fourier Embedding Aggregation (FEA) blocks to extract the decomposed geometry and lighting information, implicitly addressing the GBR ambiguity. Furthermore, we propose the Frequency-Spatial Weighted (FSW) block to enhance the fusion of frequency and local spatial information, employing an adaptively weighted function. FUPS-Net overcomes

the limitations of previous two-stage UPS methods, offering end-to-end training setups, avoiding discrete classification errors in estimating explicit lighting directions, and preventing the propagation of accumulated errors in disjointed light calibration networks and normal estimation networks. Ablation studies highlight the effectiveness of the proposed FEE, FEA, and FSW blocks. Experimental results on extensive datasets demonstrate the superior performance of FUPS-Net. We significantly improve the accuracy of the one-stage UPS and our method reaches the current best performance of the explicit lighting UPS pipelines.

Limitations and future work: Currently, FUPS-Net normalizes the intensity of photometric stereo images during training, *i.e.*, intensity calibration is not considered. While prioritizing efficiency, this design may face challenges in reconstructing objects with extreme geometric configurations or lighting conditions deviating significantly from directional lighting. Future work will focus on adapting FUPS-Net to better handle near-point light effects while maintaining its lightweight structure. Additionally, we aim to address its limitations in global context and scale-related distortions by integrating complementary techniques, such as multi-view reconstruction or scene-level priors, to improve performance in complex environments. These efforts will enhance the network's adaptability to varying lighting conditions and extend its applicability to large-scale scenes, further broadening its practical value.

ACKNOWLEDGMENT

The work was supported in part by the Ministry of Education Singapore Tier 1 grant No. RG98/24, the National Natural Science Foundation of China (62136001, 62088102). This work is partially done in NTU-ROSE Lab.

REFERENCES

[1] R. J Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 139–144, 1980.

[2] Zhenglong Zhou, Zhe Wu, and Ping Tan, "Multi-view photometric stereo with spatially varying isotropic materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1482–1489.

[3] Ufuk Sakarya, Uğur Murat Leloğlu, and Erol Tunalı, "Three-dimensional surface reconstruction for cartridge cases using photometric stereo," *Forensic Science International*, vol. 175, no. 2-3, pp. 209–217, 2008.

[4] ZhenXiong Jian, Xi Wang, XinQuan Zhang, Rong Su, MingJun Ren, and LiMin Zhu, "Task-specific near-field photometric stereo for measuring metal surface texture," *IEEE Transactions on Industrial Informatics*, 2023.

[5] Satoshi Ikehata and Kiyoharu Aizawa, "Photometric stereo using constrained bivariate regression for general isotropic surfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2179–2186.

[6] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 6, pp. 1078–1091, 2014.

[7] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Proceedings of the Asian Conference on Computer Vision, Queenstown*, 2010, pp. 703–717.

[8] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa, "Robust photometric stereo using sparse regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 318–325.

[9] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee Kenneth Wong, "Deep photometric stereo for non-lambertian surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 129–142, 2020.

[10] Satoshi Ikehata, "Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–18.

[11] E Oran Brigham and RE Morrow, "The fast fourier transform," *IEEE spectrum*, vol. 4, no. 12, pp. 63–70, 1967.

[12] Manmohan Chandraker, Jiamin Bai, and Ravi Ramamoorthi, "On differential photometric reconstruction for unknown, isotropic brdfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2941–2955, 2012.

[13] Lixiong Chen, Yinqiang Zheng, Boxin Shi, Art Subpa-Asa, and Imari Sato, "A microfacet-based reflectance model for photometric stereo with highly specular surfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3162–3170.

[14] Thoma Papadhimitri and Paolo Favaro, "A new perspective on uncalibrated photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1474–1481.

[15] Feng Lu, Imari Sato, and Yoichi Sato, "Uncalibrated photometric stereo based on elevation angle recovery from brdf symmetry of isotropic materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 168–176.

[16] Peter N Belhumeur, David J Kriegman, and Alan L Yuille, "The bas-relief ambiguity," *International Journal of Computer Vision*, vol. 35, no. 1, pp. 33–44, 1999.

[17] Boxin Shi, Yasuyuki Matsushita, Yichen Wei, Chao Xu, and Ping Tan, "Self-calibrating photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1118–1125.

[18] Thoma Papadhimitri and Paolo Favaro, "A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima," *International Journal of Computer Vision*, vol. 107, pp. 139–154, 2014.

[19] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato, "Uncalibrated photometric stereo for unknown isotropic reflectances," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1490–1497.

[20] Guanying Chen, Kai Han, and Kwan-Yee K Wong, "Ps-fcn: A flexible learning framework for photometric stereo," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–18.

[21] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong, "Self-calibrating deep photometric stereo networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8739–8747.

[22] Francesco Sarno, Suryansh Kumar, Berk Kaya, Zhiwu Huang, Vittorio Ferrari, and Luc Van Gool, "Neural architecture search for efficient uncalibrated deep photometric stereo," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 361–371.

[23] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool, "Uncalibrated neural inverse rendering for photometric stereo of general surfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3804–3814.

[24] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K Wong, and Yasuyuki Matsushita, "What is learned in deep uncalibrated photometric stereo?," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 745–762.

[25] Satoshi Ikehata, "Universal photometric stereo network using global lighting contexts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12591–12600.

[26] Satoshi Ikehata, "Scalable, detailed and mask-free universal photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13198–13207.

[27] B Shi, Z Mo, Z Wu, D Duan, SK Yeung, and P Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 271–284, 2019.

[28] Jieji Ren, Feishi Wang, Jiahao Zhang, Qian Zheng, Mingjun Ren, and Boxin Shi, "Diligent102: A photometric stereo benchmark dataset with controlled shape and material variation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12581–12590.

[29] Brian Curless and Marc Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the SIGGRAPH*, 1996, pp. 303–312.

[30] Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec, "Relighting human locomotion with flowed reflectance fields," in *Proceedings of the Eurographics conference on Rendering Techniques*, 2006, pp. 183–194.

[31] Zhuo Hui and Aswin C Sankaranarayanan, "Shape and spatially-varying reflectance estimation from virtual exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 10, pp. 2060–2073, 2016.

[32] Aaron Hertzmann and Steven M Seitz, "Example-based photometric stereo: Shape reconstruction with general, varying brdfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1254–1264, 2005.

[33] Junxuan Li, Antonio Robles-Kelly, Shaodi You, and Yasuyuki Matsushita, "Learning to minify photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7568–7576.

[34] David Honzátko, Engin Türetken, Pascal Fua, and L Andrea Dunbar, "Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo," in *Proceedings of the IEEE International Conference on 3D Vision*, 2021, pp. 394–402.

[35] Yakun Ju, Boxin Shi, Muwei Jian, Lin Qi, Junyu Dong, and Kin-Man Lam, "Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention," *International Journal of Computer Vision*, vol. 130, no. 12, pp. 3014–3034, 2022.

[36] Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, and Boxin Shi, "Gps-net: Graph-based photometric stereo network," in *Proceedings of the Advances in Neural Information Processing Systems*, 2020, p. 33.

[37] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita, "Deep photometric stereo network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 501–509.

[38] Qian Zheng, Yiming Jia, Boxin Shi, Xudong Jiang, Ling-Yu Duan, and Alex C Kot, "Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8549–8558.

[39] Huiyu Liu, Yunhui Yan, Kechen Song, and Han Yu, "Sps-net: Self-attention photometric stereo network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2020.

[40] Satoshi Ikehata, "Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism," in *Proceedings of the British Machine Vision Conference*, 2021, vol. 2, p. 11.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017.

[42] Qian Zheng, Boxin Shi, and Gang Pan, "Summary study of data-driven photometric stereo methods," *Virtual Reality and Intelligent Hardware*, vol. 2, no. 3, pp. 213–221, 2020.

[43] Yakun Ju, Kin-Man Lam, Wuyuan Xie, Huiyu Zhou, Junyu Dong, and Boxin Shi, "Deep learning methods for calibrated photometric stereo and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[44] Manmohan Krishna Chandraker, Fredrik Kahl, and David J Kriegman, "Reflections on the generalized bas-relief ambiguity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 788–795.

[45] Ondřej Drbohlav and Radim Šára, "Specularities reduce ambiguity of uncalibrated photometric stereo," in *Proceedings of the European Conference on Computer Vision*, 2002, pp. 46–60.

[46] Georghiades, "Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo," in *Proceedings of the IEEE International Conference on Computer Vision*, 2003, pp. 816–823.

[47] Zhe Wu and Ping Tan, "Calibrating photometric stereo by holistic reflectance symmetry analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1498–1505.

[48] Junxuan Li and Hongdong Li, "Self-calibrating photometric stereo by neural inverse rendering," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 166–183.

[49] Mohammed Brahimi, Bjoern Haefner, Zhenzhang Ye, Bastian Goldluecke, and Daniel Cremers, "Sparse views near light: A practical paradigm for uncalibrated point-light photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11862–11872.

[50] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian, "A fourier-based framework for domain generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14383–14392.

[51] Yanchao Yang and Stefano Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[52] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong, "Deep fourier-based exposure correction network with spatial-frequency interaction," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 163–180.

[53] Jun Yu, Peng He, and Ziqi Peng, "Fsr-net: Deep fourier network for shadow removal," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 2335–2343.

[54] Chenxi Wang, Hongjun Wu, and Zhi Jin, "Fourllie: Boosting low-light image enhancement by fourier frequency information," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 7459–7469.

[55] Chongyi Li, Chun-Le Guo, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, Chen Change Loy, et al., "Embedding fourier for ultra-high-definition low-light image enhancement," in *Proceedings of the International Conference on Learning Representations*, 2023.

[56] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 3744–3753.

[57] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen, "Hinet: Half instance normalization network for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 182–192.

[58] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao, "Frequency and spatial dual guidance for image dehazing," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 181–198.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[60] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[61] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 694–711.

[62] Micah K Johnson and Edward H Adelson, "Shape estimation in natural illumination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2553–2560.

[63] Tatsunori Taniai and Takanori Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 4857–4866.

[64] Junxuan Li and Hongdong Li, "Neural reflectance for shape recovery with shadow handling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16221–16230.

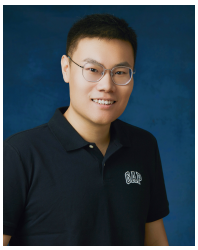
[65] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla, "Px-net: Simple and efficient pixel-wise training of photometric stereo networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 12757–12766.

[66] Ashish Tiwari and Shanmuganathan Raman, "Deepps2: Revisiting photometric stereo using two differently illuminated images," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 129–145.

[67] Ashish Tiwari and Shanmuganathan Raman, "Lerps: Lighting estimation and relighting for photometric stereo," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022, pp. 2060–2064.

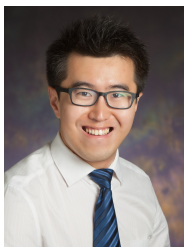
[68] Tal Simchony, Rama Chellappa, and Min Shao, "Direct analytical methods for solving poisson equations in computer vision

problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 435–446, 1990.
[69] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan, "A data-driven reflectance model," *ACM Transactions on Graphics*, pp. 759–769, 2003.
[70] Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec, "Relighting human locomotion with flowed reflectance fields," in *Proceedings of the Eurographics Symposium on Rendering*, 2006, pp. 183–194.



Yakun Ju is currently a Lecturer (Assistant Professor) with the School of Computing and Mathematical Sciences, University of Leicester, U.K. Previously, he served as a Postdoctoral Research Fellow at Nanyang Technological University and The Hong Kong Polytechnic University. Yakun earned his Bachelor's degree from Sichuan University in 2016 and his Ph.D. from Ocean University of China in 2022. His research interests span 3D reconstruction, medical image processing, underwater information perception, and computational imaging. He has authored over 50 publications in top-ranked journals and conferences, including TPAMI, TIP, TVCG, IJCV, CVPR, and NeurIPS, etc. He also contributes as an associate editor/editor board for Applied Soft Computing and Neurocomputing.

Boxin Shi received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper, Runners-Up at CVPR 2024, ICCP 2015 and selected as Best Paper candidate at ICCV 2015. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV. He is a senior member of IEEE.



Bihan Wen received the B.Eng. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign (UIUC), Champaign, IL, USA, in 2015 and 2018, respectively. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, NTU. Dr. Wen received the 2022 Early Career Teaching Excellence Award and the 2021 Inspirational Mentor for Koh Boon Hwee Award from NTU, and the 2012 Professional Engineers Board Gold Medal from Singapore. He was a recipient of the Best Paper Runner Up Award at the IEEE ICME 2020, the Best Paper Award from IEEE ICIEA 2023, and the Best Paper Award from IEEE MIPR 2023. He was ranked the World Top 2% Scientists in Artificial Intelligence from the year 2021 and 2023 consecutively by Stanford University. He was awarded the 2023 CASS VSPC Rising Star (Runner-Up). He is an Associate Editor of IEEE TCSVT. He is serving as a Guest Editor for IEEE SPM and IEEE JSTSP. His research interests include machine learning, computational imaging, computer vision, image processing, and artificial intelligence security.

Kin-Man Lam received his Associateship in Electronic Engineering with distinction from The Hong Kong Polytechnic University in 1986, his M.Sc. degree from the Department of Electrical Engineering, Imperial College, U.K., in 1987, and his Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Australia, in 1996. From 1990 to 1993 and 1996 to now, he was a lecturer, Assistant Professor, Associate Professor, and Professor at the Department of Electronic Engineering and Department of Electronic and Information Engineering of The Hong Kong Polytechnic University, respectively. Currently, he is also an Associate Dean of the Faculty of Engineering. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008, and was the Director-Student Services and the Director-Membership Services of the IEEE SPS between 2012 and 2014, and between 2015 and 2017, respectively. He was an Associate Editor of IEEE TIP between 2009 and 2014, and Digital Signal Processing between 2014 and 2018. He was also an Editor of HKIE Transactions between 2013 and 2018, and an Area Editor of the IEEE SPM between 2015 and 2017. Currently, he is the IEEE SPS VP-Membership and the Member-at-Large of APSIPA. His current research interests include image and video processing, computer vision, and human face analysis and recognition.

Xudong Jiang received the B.E. and Ph.D. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, and the PhD degree from Helmut Schmidt University, Hamburg, Germany. From 1998 to 2004, he was with the Institute for Infocomm Research, A*STAR, Singapore, as a lead scientist, and the head of the Biometrics Laboratory. He joined Nanyang Technological University (NTU), Singapore, as a faculty member, in 2004, where he served as the director of the Centre for Information Security from 2005 to 2011. He is currently a professor with the School of EEE, NTU and serves as the director of the Centre for Information Sciences and Systems of School of EEE, NTU. He holds 7 patents and has authored more than 200 papers with over 60 papers in the IEEE journals and over 30 papers in top conferences. He served as IFS TC member of the IEEE Signal Processing Society from 2015 to 2017, associate editor for IEEE Signal Processing Letter from 2014 to 2018 and associate editor for IEEE Transactions on Image Processing from 2016 to 2020. Currently, he serves as senior area editor for IEEE Transactions on Image Processing and editor-in-chief for IET Biometrics. His current research interests include image processing, pattern recognition, computer vision, machine learning, and biometrics.

Alex C. Kot has been with Nanyang Technological University, Singapore, since 1991. He was the Head of the Division of Information Engineering and the Vice Dean of Research with the School of Electrical and Electronic Engineering. Subsequently, he served as an Associate Dean for the College of Engineering for eight years. He is currently a Professor and the Director of the Rapid-Rich Object Search (ROSE) Laboratory and the NTU-PKU Joint Research Institute. He has published extensively in the areas of signal processing, biometrics, image forensics and security, and computer vision and machine learning.,He is a fellow of the Academy of Engineering, Singapore. He was elected as the IEEE Distinguished Lecturer of the Signal Processing Society and the Circuits and Systems Society. He received the Best Teacher of the Year Award. He is the co-author for several best paper awards, including ICPR, IEEE WIFS, IWDW, CVPR Precognition Workshop, and VCIP. He served at the IEEE SP Society in various capacities, such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing and the Vice-President for the IEEE Signal Processing Society. He served as an associate editor for more than ten journals, mostly for IEEE Transactions.