Revisiting Supervised Learning-Based Photometric Stereo Networks

Xiaoyao Wei, Zongrui Li, Binjie Ding, Boxin Shi, *Senior Member, IEEE,* Xudong Jiang, *Fellow, IEEE,* Gang Pan, *Senior Member, IEEE,* Yanlong Cao, and Qian Zheng, *Member, IEEE*

Abstract—Deep learning has significantly propelled the development of photometric stereo by handling the challenges posed by unknown reflectance and global illumination effects. However, how supervised learning-based photometric stereo networks resolve these challenges remains to be elucidated. In this paper, we aim to reveal how existing methods address these challenges by revisiting their deep features, deep feature encoding strategies, and network architectures. Based on the insights gained from our analysis, we propose ESSENCE-Net, which effectively encodes deep shading features with an *easy-first-encoding* strategy, enhances shading features with shading supervision, and accurately decodes normal with spatial context-aware attention. The experimental results verify that the proposed method outperforms state-of-the-art methods on three benchmark datasets, whether with dense or sparse inputs. The code is available at https://github.com/wxy-zju/ESSENCE-Net.

Index Terms—Photometric stereo, Deep features, Shading supervision, Easy-first-encoding, Spatial context-aware attention

1 Introduction

PHOTOMETRIC stereo (PS) aims to recover the surface normal from images captured under varying light directions at a fixed viewpoint [1]. With the capability to recover pixel-level fine details, PS holds promising applications in many fields, such as industrial inspection [2], cultural heritage preservation [3], and lunar surface reconstruction [4]. However, classical PS [1] assumes Lambertian surfaces without global illumination effects, deviating from real-world scenarios. Conventional methods relax the assumption by modeling surface reflectance explicitly through simplified analytic models [5], [6], [7], [8] and addressing global illumination effects through outlier rejection [9], [10], [11], [12]. However, reflectance and global illumination effects are often unknown and vary across different subjects or even the same object, which are two main challenges in PS.

In recent years, deep learning has significantly improved the accuracy of normal estimation [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], facilitating modeling reflectance of various materials and even relaxing constraints of physical light models (*e.g.*, universal light [29], [31], [32], sunlight [33], screen light [34]). Additionally, it enables real-time reconstruction with event cameras [35] or achieves high accuracy with reinforcement learning [36] or inverse rendering (with 2 images) [37]. Compared to traditional methods, supervised learning exhibits superior performance in PS because of its nonlinear fitting and generalization ability from explicit labels [38], especially on large datasets according to the scaling law [39]. Consequently, investigating the essence of supervised learning-based photometric stereo networks (SL-PSNs) in addressing the main challenges of PS becomes crucial.

Previous SL-PSNs can be categorized into per-pixel [14], [15], [16], [17], [18], all-pixel [19], [20], [21], [22], [23], [24], and hybrid methods [25], [26], [27], [28] based on input types (1D-profile[1], 2D-image[1], and both, respectively), as proposed in [41], [42]. Regardless of the inputs, SL-PSNs consist of three stages: *observation-light alignment*, *light-aware deep feature encoding*, and *light-free deep feature decoding* due to the unstructured input (*i.e.*, order-agnostic and arbitrary number of images) [14], as shown in Fig. 1. In *observation-light alignment*, per-pixel methods project 1D-profile onto an observation map indexed by light directions inefficiently, while all-pixel and hybrid methods align 2D-image and light map[2] through simple concatenation efficiently. Then, in *light-aware deep feature encoder*, per-pixel (all-pixel) methods extract deep features from aligned 1D-profiles (2D-images) using shared weights CNNs (siamese networks), while hybrid methods combine 1D-profiles and 2D-images to facilitate deep feature extraction. Finally, in *light-free deep feature decoder*, light-aware deep features are fused to eliminate light information through weighted summation (per-pixel), max-pooling (all-pixel and hybrid), or transformer decoder

- *Xiaoyao Wei, Binjie Ding, and Yanlong Cao are with the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, Hangzhou, Zhejiang 310058, China, and also with Zhejiang Key Laboratory of Advanced Equipment Manufacturing and Measurement Technology, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: {w_xy, dingbinj, sdcaoyl}@zju.edu.cn.*
- *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, 100871, China. E-mail: shiboxin@pku.edu.cn.*
- *Zongrui Li and Xudong Jiang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798. E-mail: {zongrui001, exdjiang}@ntu.edu.sg.*
- *Qian Zheng and Gang Pan are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China, and also with the State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, Zhejiang 310027, China. E-mail: {qianzheng, gpan} @zju.edu.cn.*

Corresponding authors: Zongrui Li, Yanlong Cao and Qian Zheng

1. 1D-profile is a set of intensity variations at one point under different lights [40]. 2D-image is the observed image under a specific light.

2. A light map is formed by replicating the 3-vector light direction to match the spatial dimension of the observed image.

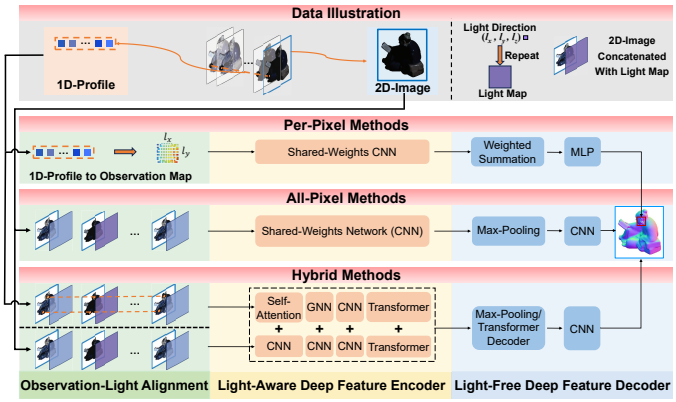


Fig. 1. Illustration of two-stage feature extraction for supervised learning-based photometric stereo networks. Per-pixel, all-pixel, and hybrid methods are categorized based on input types, as proposed in [41], [42].

(hybrid). Normals are then decoded from light-free deep features using a multi-layer perceptron (MLP) (per-pixel) or CNNs (all-pixel and hybrid).

However, despite the excellent performance of SL-PSNs, how challenges of unknown reflectance and global illumination effects are addressed during the network's feedforward process remains to be elucidated. Despite some studies [19], [20] indicating a correlation between deep features and shading, they lack quantitative studies and fail to leverage their studies to facilitate resolving these two challenges.

To further investigate how challenges of unknown reflectance and global illumination effects are resolved by SL-PSNs, we **revisit SL-PSNs in aspects of deep feature, deep feature encoding, and network architectures** and attempt to answer the following three fundamental questions. 1) What is the desired deep feature to be learned during neural network optimization? 2) How to facilitate resolving the challenges of PS (*i.e.*, unknown reflectance and global illumination effects) during deep feature encoding? 3) What is the desired network architecture that facilitates extracting desired deep features and resolving PS challenges? The contributions are summarized as follows:

To answer the first question, we revisit the deep features learned by conventional SL-PSNs and demonstrate that these approaches are dedicated to extracting features highly correlated to shading in the stage of *light-aware deep feature encoding*. Based on this observation, we propose shading supervision loss to facilitate feature extraction during this stage.

To answer the second question, we reveal that per-pixel/all-pixel methods are more adapt at addressing the challenge of unknown reflectance/global illumination effects, based on which the difficulty of addressing these challenges for deep feature encoding is analyzed. We show that the difficulty is relevant to the information of inputs (*e.g.*, image resolution, number, and non-uniform material). Inspired by *easy-first-generation/least-to-most* principle applied in diffusion models [43], [44]/large language models [45], we propose an *easy-first-encoding* strategy to develop *light-aware deep feature encoder* in hybrid methods, facilitating resolving unknown reflectance and global illumination effects.

To answer the third question, we provide an analysis of clues utilized for resolving PS challenges, based on which a discussion about how network architectures in existing SL-PSNs facilitate resolving these challenges for *light-aware deep feature encoding* is provided. We further propose spatial context-aware attention for *light-free deep feature decoding*.

Combining above insights, we propose **ESSENCE-Net**, which **E**ncodes light-aware deep features with **S**hading **S**upverision and **E**asy-first-encoding strategy, and decodes light-free features i**N** a spatial **C**ontext-aware attention scheme, for surface normal **E**stimation. ESSENCE-Net is validated to achieve state-of-the-art performance on the *DiLiGenT* [46], *DiLiGenT10*^{2} [47], and *DiLiGenT-II* [48] benchmark datasets, both in dense and sparse settings.

2 Related Work

This section reviews SL-PSNs based on the three-stage feature extraction. Tab. 1 compares the differences between representative SL-PSNs and the proposed ESSENCE-Net. We follow the traditional setup of PS, which assumes orthographic projection and directional lights. Other setups are beyond the scope of this paper; readers may refer to [49], [50], [51] for near light, [33] for sunlight, [34] for screen light, [29], [31], [32] for universal light, and [52] for uncalibrated light. For unsupervised learning-based photometric stereo networks, readers can refer to [53], [54], [55], [56], [57].

2.1 Observation-Light Alignment

In PS, varying lights provide clues for normal recovery. Evidence [19], [20] has shown that incorporating light information with the observed images can enhance the accuracy of normal recovery. Per-pixel methods [14], [15], [16], [17], [18] map the observed pixel values of a point under varying lights into an observation map, where the position represents the light direction. All-pixel [19], [20], [21], [22], [23], [24] and hybrid [25], [26], [27], [28] methods expand the 3-vector light direction to match the size of the image, forming image-light pairs through simple concatenation. While the pixel-wise generated observation map increases the input size and reduces the efficiency of normal estimation. The simple concatenation of image and light allows the recovery of the entire normal map in an end-to-end manner, leading to higher efficiency. Therefore, the proposed ESSENCE-Net adopts the simple image-light concatenation approach for efficient normal estimation.

2.2 Light-Aware Deep Feature Encoding

In *light-aware deep feature encoder*, the desired features for normal recovery are extracted. Existing SL-PSNs primarily focus on this stage to address the challenges posed by unknown reflectance and global illumination effects.

Per-pixel methods feed observation maps into CNNs to encode deep features from 1D-profiles, achieving excellent results [14], [15], [17]. However, the performance significantly degrades in sparse setups. Networks specially designed for sparse setups are proposed, such as SPLINE-Net [18], which generates dense observation maps by light interpolation, and LMPS [16], which applies a connection table to select the most useful lights. Since per-pixel methods

TABLE 1
A Summary of the Proposed ESSENCE-Net and Representative SL-PSNs.

Method	Observation-Light Alignment		Light-Aware Deep Feature Encoder			Light-Free Deep Feature Decoder		Supervision	Performance on *DiLiGenT*
	Scheme	Speed	Deep Features	Encoding Scheme	Architecture	Light-Aware Feature Fusion	Architecture		
CNN-PS [14]	Observation Map	Slow	N.C.	Unmixed (1D-Profile)	CNN	Weighted Summation	MLP	Normal	7.21°
LMPS [16]	Observation Map	Slow	N.C.	Unmixed (1D-Profile)	CNN	Weighted Summation	MLP	Normal+Conn.tab.	8.43°
SPLINE-Net [18]	Observation Map	Slow	N.C.	Unmixed (1D-Profile)	CNN	Weighted Summation	MLP	Normal+Sym.+Asym.	9.63°
PX-Net [49]	Observation Map	Slow	N.C.	Unmixed (1D-Profile)	CNN	Weighted Summation	MLP	Normal	6.17°
PS-FCN [19]	Simple Concatenation	Fast	Shading	Unmixed (2D-Image)	CNN	Max-Pooling	CNN	Normal	8.13°
Attention-PSN [21]	Simple Concatenation	Fast	N.C.	Unmixed (2D-Image)	CNN	Max-Pooling	CNN	Normal	7.92°
SPS-Net [26]	Simple Concatenation	Fast	N.C.	Mixed (Per-All-Alternate)	Self-Attention+CNN	Max-Pooling	CNN	Normal	7.60°
GPS-Net [28]	Simple Concatenation	Fast	N.C.	Mixed (Per-First-Encoding)	GNN+CNN	Max-Pooling	CNN	Normal	7.81°
MT-PS-CNN [25]	Simple Concatenation	Fast	N.C.	Mixed (Per-First-Encoding)	CNN+CNN	Max-Pooling	CNN	Normal	7.56°
PS-Transfromer [27]	Simple Concatenation	Fast	N.C.	Mixed (Per-All-Parallel)	Transformer+Transformer	Transformer Decoder	CNN	Normal	7.66°
ESSENCE-Net	Simple Concatenation	Fast	Shading	Mixed (**Easy-First-Encoding**)	**Transformer+CNN**	Max-Pooling	**Transformer**	**Normal+Shading**	5.69°

Notes: The performance on *DiLiGenT* [46] of PS-Transformer [27] is evaluated with 10 images, while the others are with 96 images. GPS-Net [28] places the feature extraction of 2D-images in *light-free deep feature decoder*. 'N.C.' represents Not Considered. 'Encoding Scheme' includes unmixed (1D-Profile/2D-Image) and mixed schemes (Per-First-Encoding/All-First-Encoding/Per-All-Alternate/Per-All-Parallel). 'Per-First-Encoding' means first using the per-pixel method to extract deep features from 1D-profiles, then integrating the all-pixel method to enhance shading features from 2D-images, whereas "ALL-First-Encoding" operates conversely. 'Per-All-Alternate' alternates between the two operations, and 'Per-All-Parallel' means performing both in parallel. 'Conn.tab.' means the regularization loss over the connection table proposed in [16]. 'Sym.' and 'Asym.' mean symmetric and asymmetric loss proposed in [18].

ignore spatial information, they are sensitive to global illumination effects, requiring specific strategies (*e.g.*, shadow map [17], occlusion layer [16], and asymmetric loss [18]).

All-pixel methods [19], [20], [21], [22], [23], [24] primarily exploit information from 2D-images to obtain deep features. These methods feed aligned image-light pairs into a shared-weight siamese network, extracting deep features from shape clues with the help of the cross-correlation principle of CNN [58]. However, the lack of explicit model reflectance makes all-pixel methods sensitive to non-uniform materials.

Hybrid methods [25], [26], [27], [28] adopt a mixed scheme to encode deep features from 1D-profiles and 2D-images, thus solving unknown reflectance and global illumination effects. For the solving scheme, GPS-Net [28] and MT-PS-CNN [25] use a sequential per-pixel followed by all-pixel method, PS-Transformer [27] performs both in parallel, and SPS-Net [26] alternates between the two operations. Architectures such as CNN [25], graph neural network (GNN) [28], transformer [27], and self-attention [26] are utilized for feature extraction from 1D-profiles or 2D-images.

Despite the advancements of SL-PSNs, they have not explicitly revealed the deep features encoded at this stage. Although some studies [19], [20] indicate a correlation between deep features and shading, their analysis focuses on light-free features, where shading is related to light direction. Moreover, their analysis lacks quantification validation.

The proposed ESSENCE-Net also adopts a mixed scheme for encoding light-aware deep features, but it differs from all previous methods in three aspects. 1) We analyze the essence of SL-PSNs in addressing the challenges of PS and validate that the learning involves removing unknown reflectance and global illumination effects, essentially purifying features highly correlated to shading. Based on this observation, ESSENCE-Net employs shading supervision to enhance shading features. 2) ESSENCE-Net employs an input-aware mixed scheme (*easy-first-encoding*) based on analyzing the difficulty in resolving unknown reflectance and global illumination effects across various cases, distinct from fixed strategies in [25], [26], [27], [28]. 3) We present an analysis of network architectures for resolving unknown reflectance and global illumination effects to encode deep features.

2.3 Light-Free Deep Feature Decoding

In *light-free deep feature decoder*, normals are decoded from the encoded deep features, which receive less attention in existing research. Initially, observations under K lights are fused to eliminate light-related features, where weighted summation are used in per-pixel methods [14], [15], [16], [17], [18], max-pooling [19], [20], [21], [22], [24], [30] (or multi-scale max-pooling [23]) is used in all-pixel methods, and transformer decoder [27] or max-pooling [25], [26], [28] is used in hybrid methods. Subsequently, per-pixel methods employ an MLP to recover the 3-vector normal independently, while all-pixel or hybrid methods use CNNs to recover the 3-channel normal map spatially. The proposed ESSENCE-Net also utilizes max-pooling to fuse deep features. However, unlike per-pixel methods, we consider shape similarity for normal recovery. Unlike all-pixel and hybrid methods, we propose spatial context-aware attention, as commonly used CNNs can only leverage local shape similarity due to the limited effective receptive field [59].

3 Revisiting Supervised Learning-based Photometric Stereo Networks (SL-PSNs)

This section revisits SL-PSNs in tackling two challenges: unknown reflectance and global illumination effects. First, we show that SL-PSNs inherently learn features highly correlated to shading to remove unknown reflectance and global illumination effects. Next, we analyze the difficulty of addressing those two challenges from 1D-profiles or 2D-images, proposing an *easy-first-encoding* strategy for optimized shading features. Finally, we revisit the network architectures to address those two challenges for encoding shading features and propose spatial context-aware attention to decode normals from them. These analyses instruct the network design in Section 4.

3.1 Revisiting Deep Feature

In this section, we show that per-pixel (all-pixel) methods primarily learn deep features highly correlated to shading from reflectance (shape) clues, *i.e.*, addressing challenges of unknown reflectance and global illumination effects by removing them. These analyses guide the network design in Section 4.1.

3.1.1 Analysis

Given a point with normal $\boldsymbol{n} \in \mathbb{R}^3$ illuminated by light with direction $\boldsymbol{l} \in \mathbb{R}^3$, the observed intensity I is given by

$$I = \rho(\boldsymbol{n}, \boldsymbol{l}, \boldsymbol{v})\max(\boldsymbol{n}^{\mathrm{T}}\boldsymbol{l}, 0) + \eta, \tag{1}$$

where ρ is the reflectance, $\max(\boldsymbol{n}^{\mathrm{T}}\boldsymbol{l}, 0)$ is shading with attached shadows, η is global illumination effects (cast shadows and inter-reflections), $\boldsymbol{v} \in \mathbb{R}^3$ is the view direction.

It is obvious that I is influenced by ρ, $\max(\boldsymbol{n}^{\mathrm{T}}\boldsymbol{l}, 0)$, and η. Traditional methods model ρ and η by simplified reflectance models [5], [6], [7], [8] and outlier rejection [9], [10], [11], [12], allowing shading to be isolated for normal recovery. For SL-PSNs, we posit that deep features learned from 1D-profiles or 2D-images convey less information relevant to ρ or η but remove them to describe $\max(\boldsymbol{n}^{\mathrm{T}}\boldsymbol{l}, 0)$, *i.e.*, shading.

Encoding shading. SL-PSNs aim to estimate the normal $\boldsymbol{n}$, and the deep features inherently include light $\boldsymbol{l}$ (shown in Fig. 1) in *light-aware deep feature encoder*. Thus, it is reasonable to assume that shading $\max(\boldsymbol{n}^{\mathrm{T}}\boldsymbol{l}, 0)$ (linearly related to normal) constitutes the major part of the deep features.

Resolving reflectance. Although all-pixel methods fuse deep features with max-pooling for deep feature decoding, they lack the utilization of intensity variations under different lights (absent in 2D-images) for deep feature encoding, thereby failing to model reflectance [41], [42]. In contrast, per-pixel methods utilize reflectance clues from intensity variations in 1D-profiles to encode deep features [41], [42]. However, the highly nonlinear relationship between reflectance and normal [40] makes it challenging to recover normal with several dense layers [14], [15], [16], [17], [18].

Resolving global illumination effects. Although per-pixel methods consider global illumination effects (abrupt intensity changes in the observation map [14]) through shadow map [17], occlusion layer [16], or asymmetric loss [18], they lack the utilization of shape clues that contribute to these effects (absent in 1D-profiles). In contrast, all-pixel methods complement global illumination effects by leveraging shape clues from 2D-images [41], [42]. However, recovering the entire normal map from them is difficult due to the sparse distribution and vague relationship between global illumination effects and normals.

Previous studies also support our hypothesis. Chen *et al.* [19], [20] found a correlation between fused features and shading (belonging to all-pixel methods). However, such correlation can only be qualitatively verified, not quantitatively, because they focus on light-free deep features, while shading is related to the light direction. Moreover, the features extracted by per-pixel methods remain unexplored. In contrast, we fill these gaps in the following section.

3.1.2 Validation

To validate the hypothesis proposed in Section 3.1.1, we conduct quantitative correlation and qualitative visualization analysis of deep features extracted by SL-PSNs on *DiLiGenT* dataset [46]. We select the most typical per-pixel (CNN-PS [14]) and all-pixel (PS-FCN [19]) methods, whose concept has been adopted by many state-of-the-art methods [15], [16], [17], [18], [20], [52], [60]. Particularly, we consider the corresponding types of deep features, *i.e.*, the average and max value at channel dimension in different layers for CNN-PS [14] and PS-FCN [19], respectively[3]. Following results are based on the released models by authors of [14] and [19].

Correlation analysis. We use the mean absolute error between the deep features and ground truth shading (shading error) for quantitative analysis. According to Fig. 2, the shading error for both CNN-PS [14] and PS-FCN [19] decreases as the network goes deeper, validating our hypothesis that the encoded features (especially features in the deep layer) have a strong correlation with shading. Moreover, we find a strong correlation between the shading error of the deepest layer (Feature 5) and the normal estimation error of L2 method [1] regarding different objects[4]. Since the L2 method directly recovers normals from shading, the normal estimation error inversely reflects the deviation from shading. Therefore, such a strong correlation manifests that the shading error we compute is highly correlated with the actual deviation from the ground truth shading, which enhances the validity of our quantitative verification.

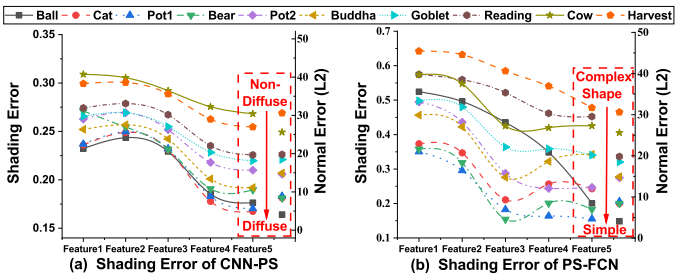


Fig. 2. Quantitative analysis of the relationship between extracted features by SL-PSNs and shading on 10 objects of *DiLiGenT* dataset [46]. Features 1 to 5 refer to the features extracted by the first five CNN layers in *light-aware deep feature encoder*. (a) Shading error of CNN-PS [14]. Diffuse materials are easier to handle, whereas complex shapes are more difficult. (b) Shading error of PS-FCN [19]. Simple shapes are easier to handle, whereas complex shapes are more difficult.

Besides, we find that the shading error of CNN-PS [14] is more reflectance-dependent, showing larger errors for non-diffuse surfaces and smaller errors for diffuse ones, as it exploits reflectance clues, while PS-FCN [19] is more shape-dependent, exhibiting larger errors for complex shapes and smaller errors for simple ones, as it exploits shape clues.

Visualization analysis. Fig. 3 illustrates the visualized features in different layers of PS-FCN [19] and CNN-PS [14][5]. For CNN-PS, we select a few surface points to visualize the corresponding features [14]. For PS-FCN [19], we show the shading features under specific light directions. It is obvious that as the network goes deeper, the features of both methods become more like shading. Additionally, we observe that per-pixel methods still assign non-zero values for regions under the attached shadows, as shown in Fig. 3. We speculate that the CNN may assign unrelated values, as the attached shadows contain less useful information. However, for all-pixel methods, attached shadows tend to be relatively preserved due to a feature extraction on entire images with shape clues, as shown in Fig. 3.

We analytically and experimentally show that SL-PSNs inherently learn features highly correlated to shading, *i.e.*, removing unknown reflectance and global illumination effects, where per-pixel methods focus more on modeling reflectance from 1D-profiles, all-pixel methods primarily

3. More details on how to calculate the deep features can be found in the supplementary material.

4. The correlation coefficient is 0.935 (0.912) for CNN-PS [14] (PS-FCN [19]), more details can be found in the supplementary material.

5. More randomly sampled qualitative results can be found in the supplementary material.

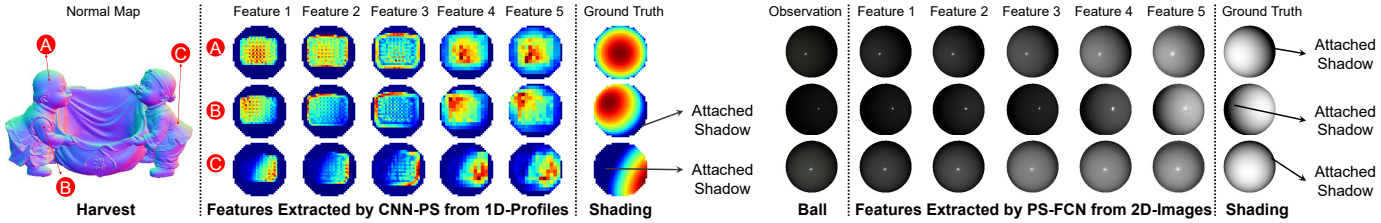


Fig. 3. Visualization of the deep features extracted from CNN-PS [14] and PS-FCN [19]. Feature 1 to Feature 5 refer to the features extracted by the first five CNN layers in *light-aware deep feature encoder*. The attached shadow is shown in the ground truth shading, either in the blue region (CNN-PS [14]) or the black region (PS-FCN [19]).

utilizing shape clues from 2D-images. These insights guide the network design in Section 4.1 (*e.g.*, shading supervision) and form the basis for the subsequent analysis.

3.2 Revisiting Deep Feature Encoding

In this section, we first analyze the difficulty of addressing PS challenges (*i.e.*, unknown reflectance and global illumination effects) on shading encoding from 1D-profiles or 2D-images and demonstrate that such difficulty is correlated with information availability. We then propose an *easy-first-encoding* strategy to facilitate resolving these two challenges and encode more accurate shading features. These analyses form the basis of the network design in Section 4.2.

3.2.1 Analysis of Per-Pixel or All-Pixel Methods

In PS, the input data includes information from three dimensions, where 2D-images (*i.e.*, the height and width dimensions in the images) and 1D-profiles (*i.e.*, the light dimension) are utilized by SL-PSNs to address unknown reflectance and global illumination effects.

Per-pixel methods. Per-pixel methods [14], [15], [16], [17], [18] primarily utilize the information in 1D-profiles. The observation I_i^{1DP} under i-th light can be described as:

$$I_i^{1DP} = \rho\left(\boldsymbol{n}, \boldsymbol{l}_i, \boldsymbol{v}\right) \max\left(\boldsymbol{n}^{\mathrm{T}}\boldsymbol{l}_i, 0\right) + \eta_i. \tag{2}$$

where $1 \leq i \leq K$, and K represents the number of lights. The variations in 1D-profile are primarily caused by changes in known light directions. Thus, per-pixel methods are more adept at modeling reflectance ρ. As a result, the deep features extracted from 1D-profiles are mainly affected by image number: the network struggles to exploit reflectance clues to encode shading features given limited input images (*e.g.*, CNN-PS [14] performs poorly under sparse setup); reflectance clues can be efficiently exploited given adequate input images to encode deep features (*e.g.*, various analytic models [5], [6], [7], [8] can be fitted).

All-pixel methods. All-pixel methods [19], [20], [21], [22], [23], [24] primarily utilize the variations in 2D-images. The observation I_j^{2DI} of the j-th pixel in an image with height H and width W under light direction $\boldsymbol{l}$ can be described as:

$$I_j^{2DI} = \rho\left(\boldsymbol{n}_j, \boldsymbol{l}, \boldsymbol{v}\right) \max\left(\boldsymbol{n}_j^{\mathrm{T}}\boldsymbol{l}, \boldsymbol{0}\right) + \eta_j. \tag{3}$$

where $1 \leq j \leq HW$. The variations in 2D-image are primarily caused by changes in shape $\boldsymbol{n}$. Thus, all-pixel methods rely on shape clues and the cross-correlation principle of CNN [58] to encode deep features. As a result, deep features extracted from 2D-images is affected by the image resolution: while limited neighborhood information makes shading extraction challenging given low resolution, more accurate shading features can be extracted given higher resolution. However, all-pixel methods struggle to handle spatially varying bidirectional reflectance distribution function (SVBRDF) surfaces [20], [42] as non-uniform materials harm CNN's ability to extract cross-correlated information.

In summary, the number and resolution of images affect the performance of per-pixel and all-pixel methods, respectively; SVBRDF surfaces also affect all-pixel methods.

Case	I		II		III		IV	
Factors	Image Number/Resolution						SVBRDF	
Example	Low Resolution		Small Image Number		H K W		Non-Uniform	
Relative Difficulty	1D-Profile	2D-Image	1D-Profile	2D-Image	1D-Profile	2D-Image	1D-Profile	2D-Image
	Easy	Difficult	Difficult	Easy	Easy	Difficult	Easy	Difficult
Method	Per-First-Encoding		All-First-Encoding		Per-First-Encoding		Per-First-Encoding	

Fig. 4. *Easy-first-encoding* strategy on different cases.

3.2.2 Analysis of the Hybrid Methods

Hybrid methods combine per-pixel and all-pixel methods to handle unknown reflectance and global illumination effects in *light-aware deep feature encoder*. The mixed encoding scheme for deep features includes per-first-encoding[6], all-first-encoding[6], per-all-alternate[6], and per-all-parallel[6]. However, previous hybrid methods employ different and fixed strategies (*e.g.*, per-first-encoding [25], [28], per-all-alternate [26], per-all-parallel [27]), failing to achieve optimal performance and lacking generalizability. In fact, deep networks typically learn more accessible features first and combine them to address complex problems later [38], [61], *e.g.*, *easy-first-generation* strategy applied in diffusion models [43], [44], *least-to-most* strategy applied in large language models [45]. Inspired by these, we propose an *easy-first-encoding* approach, *i.e.*, to encode deep features from 1D-profiles or 2D-images first while the other next in a sequential mixed method, based on which has sufficient information to improve SL-PSNs' performance. Based on the difficulty analyzed in Section 3.2.1 and insight of *easy-first-encoding* strategy, we categorized the cases according to relative difficulty, as shown in Fig. 4.

Case I: low resolution of the images results in limited information in 2D-images for encoding shading features,

6. For a detailed explanation, see the notes in Tab. 1.

while 1D-profiles provide relatively sufficient information. Therefore, feature extraction should be performed in the order of per-first-encoding.

Case II: insufficient images result in limited information in 1D-profiles for modeling reflectance, while 2D-images provide relatively sufficient information. Thus, all-first-encoding is preferred.

Case III: sufficient images with high resolution result in sufficient information in both 1D-profiles and 2D-images. However, Eq. (2) shows that resolving $\boldsymbol{n}$ from several I_i^{1DP} is less difficult than resolving $\boldsymbol{n_j}$ from several I_j^{2DI}. Because across different I_i^{1DP}, there is only one unknown $\boldsymbol{n}$, whereas across different I_j^{2DI}, there is several unknowns $\boldsymbol{n_j}$. In this case, per-first-encoding strategy will perform better.

Case IV: non-uniform materials harm the performance to extract shading features from 2D-images, while less affect extraction from 1D-profiles. Therefore, per-first-encoding should be implemented.

In practice, we find normalization strategy [14], [20], [24], [28] may affect the performance of *easy-first-encoding* strategy. While it benefits shading encoding by mitigating the effect of Lambertian reflectance, which improves the performance of normal estimation [20], its performance is less effective on anisotropic material or complex-shaped surfaces. We conduct more experimental validation on the effectiveness of the *easy-first-encoding* strategy across **Case I** to **Case IV** and the role of normalization in Section 5.2.

3.3 Revisiting Network Architectures

In this section, we analyze suitable network architectures for resolving unknown reflectance and global illumination effects to encode light-aware deep features and propose spatial context-aware attention to decode the normal map from light-free deep features (detailed in Section 4.3). These analyses guide the network design in Sections 4.2 and 4.3.

3.3.1 Architectures for Deep Feature Encoding

An effective design for a deep feature encoder should be tailored to the characteristics of different types of input data. As discussed in Section 3.1, 1D-profiles capture intensity variations caused by the reflectance at specific points, while 2D-images contain global illumination effects under a given light direction. To address PS challenges in these two types of data through feature encoding, we consider transformer as a more suitable network structure for 1D-profiles, and CNNs for 2D-images. In the following part of this section, we analyze the rationale behind this choice with respect to the distinct properties of each data type.

1D-profiles. An encoder for 1D-profiles must account for the varying reflectance of objects to effectively extract shading features. However, the sequential nature of 1D-profiles requires the encoder to be order-agnostic, ensuring that reordering elements in the sequence does not significantly affect performance. Furthermore, 1D-profiles exhibit randomness across different points, and the amount of beneficial information for shading feature extraction (such as Lambertian-dominant reflectance or low-frequency components [32], [40]) lacks consistency. This inconsistency is validated through statistical analysis of low-frequency information content on the *DiLiGenT* dataset [46], as shown in Fig. 5.

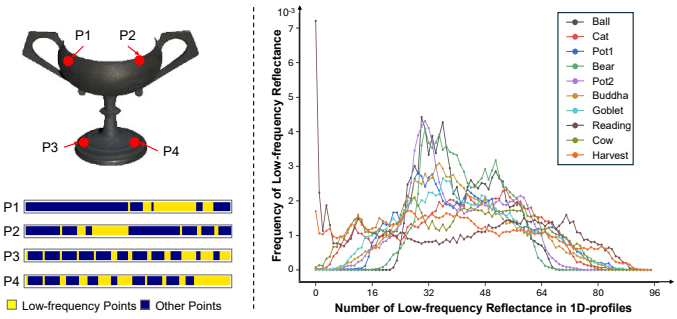


Fig. 5. Visualization of low-frequency reflectance components on 1D-profiles and statistics on frequency of low-frequency reflectance (point's reflectance that falls between 5 and 50 percentage of all points' reflectance) with respect to the number of low-frequency reflectance in 1D-profiles.

Specifically, we establish upper and lower intensity bounds (5th and 50th percentiles of all pixel values) to identify points within this intensity range. A one-hot vector is then formed, where a value of 1 represents low-frequency points, and 0 represents the others. The number of 1 in the vector reflects the amount of beneficial information in the 1D-profiles. Statistical results show that nearly all objects exhibit a flattened distribution in the amount of this information, highlighting the inherent complexity of 1D-profiles.

While CNNs [14], [15], [16], [17], [25], GNNs [28], self-attention [26], and transformer [27] are potential choices for encoding 1D-profiles, we consider transformers to be the most suitable option given these challenges. Specifically, CNNs are inherently limited in handling spatial information in 1D sequential data due to their restricted receptive fields and sensitivity to light-order variations. Converting 1D-profiles into a 2D observation map may partially address these limitations [14], but this approach is constrained in sparse setups and less adaptable to hybrid network structures. GNNs and self-attention offer global feature utilization and order-agnostic, but the performance of GNNs is sensitive to node distance [62], and self-attention may converge exponentially to a rank-1 matrix with depth [63]. In contrast, transformers overcome these issues by leveraging global reflectance patterns to extract deep shading features from 1D-profiles while maintaining order-agnostic. A performance comparison in Tab. 2 (Section 5) validates the superiority of transformers for this task. Considering the improvement, the extra computational consumption introduced by the transformer can be ignored.

2D-images. An encoder for 2D-images should address global illumination effects, with cast shadows being a key contributor, to extract shading features. In PS, these effects are often clustered in specific object regions and exhibit limited correlation with distant regions, validated by a similarity analysis on shadow patterns cast by different lights. Specifically, we represent shadow patterns using a one-hot vector for each surface point, where low pixel values ($<10^{-6}$) are marked as 1 (shadow) and others as 0. We then compute the cosine similarity between pairs of points. As shown in Fig. 6, correlated shadow patterns are primarily distributed within a small distance. Further statistical analysis of shadow pattern similarity versus distance across

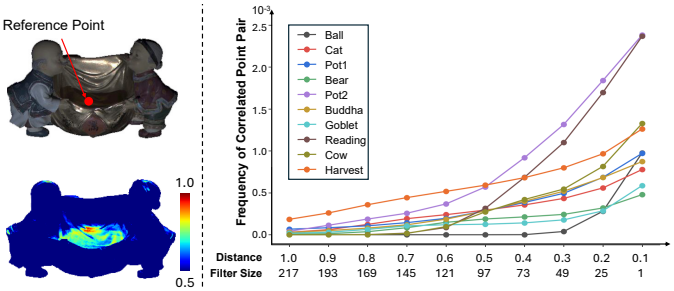


Fig. 6. Visualization of correlations between shadow patterns of reference point and other points on object's surface, as well as statistics on frequency of correlated point pairs with respect to the distance and corresponding referenced filter size.

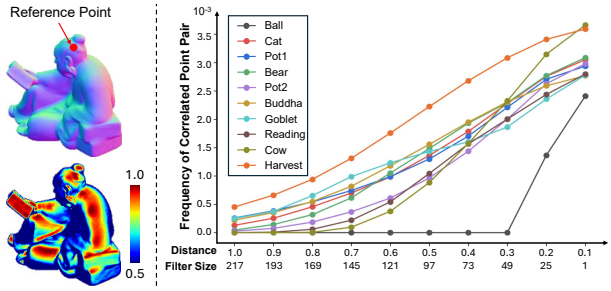


Fig. 7. Visualization of correlations between normals of reference point and other points on object's surface and statistics on frequency of correlated point pairs with respect to the distance and corresponding referenced filter size.

all point pairs (Fig. 6) confirms that useful information for addressing global illumination effects is concentrated in the local region around each point. This analysis indicates that utilizing global information for normal estimation provides limited improvement, as the correlation of global patterns diminishes with distance (Tab. 2, comparison between IDs 11 and 12). Additionally, leveraging global information significantly increases training and inference costs. In contrast, CNNs [19], [24] are a more suitable choice for 2D-image encoders when computational resources are limited, due to their strong ability to exploit local information while maintaining significantly lower computational costs.

3.3.2 Architectures for Deep Feature Decoding

The goal of feature decoding is to extract surface normal from shading features under K lights. Since shading features contain both normal- and light-related information, it is crucial to first fuse the features across K lights to remove light-related components, and then decode the normals from normal-related information. Per-pixel methods employ fully connected layers for feature fusion and decoding, followed by an L2-normalization layer for pixel-wise normal estimation [14], [15], [16], [17], [18], while all-pixel and hybrid methods use max-pooling [19], [20], [24], [25], [26] or transformer [27] for feature fusion, followed by simple CNNs with L2 normalization [19], [20], [24], [25], [26] or MLP [29] for normal estimation. Despite all-pixel or hybrid methods showing advantages in exploiting shape clues like local similarity and continuity of surface normals distribution, it is also restricted by the effective receptive fields [59]. Incorporating global information is important for more accurate normal extraction, since normals exhibit global similarity and continuity. This implies that both distant and nearby points' deep features can provide valuable information for estimating normal at a specific point.

We illustrate the global similarity and continuity in Fig. 7. Specifically, we calculate the cosine similarity between pairs of normal vectors and plot the number of correlated point pairs (those with a cosine similarity greater than 0.8) beyond varying distances in image coordinates. Additionally, we record the reference filter size required to include these points' information. As shown in Fig. 7, many points exhibit high correlation even when they are far apart, while revealing shape-related clues. Generally, the plotted curve is more convex compared to the previous shadow curve in Fig. 6. Leveraging such correlations would require either very large filter sizes or deep CNN architectures, which are less effective compared to transformers. This is validated by the comparison between IDs 1 and 6 in Tab. 2. Based on these insights, we propose spatial context-aware attention for *light-free deep feature decoding*. The analyses above form the foundation of the model design detailed in Section 4.3, with further experimental validation provided in Section 5.3.3.

4 Deep Shading Neural Network

In this section, we introduce the proposed ESSENCE-Net based on the insights of revisiting SL-PSNs in Section 3.

The overall architecture of ESSENCE-Net is illustrated in Fig. 8, which comprises two parts: *light-aware shading feature encoder*[7] and *light-free shading feature decoder*. We first align images (normalized by maximum [14]) with light directions through simple concatenation, allowing for rapid normal estimation in an end-to-end manner. Specifically, based on the analysis in Section 2.1, we replicate the 3-vector light direction into a $3 \times H \times W$ light map and concatenate it with the $3 \times H \times W$ image to form a $6 \times H \times W$ image-light pair. Subsequently, all aligned image-light pairs under K light directions are stacked to form a $6 \times K \times H \times W$ input. The stacked input is fed into *light-aware shading feature encoder* to obtain deep shading features, which are then processed by *light-free shading feature decoder* to generate accurate normal maps.

The proposed ESSENCE-Net is trained by the guidance of shading features in *light-aware shading feature encoder*, as described in Section 4.1. We develop the blocks in encoder according to the *easy-first-encoding* strategy, described in Section 4.2. Additionally, ESSENCE-Net is designed to decode accurate normal maps in *light-free shading feature decoder* with spatial context-aware attention, as detailed in Section 4.3.

4.1 Deep Feature Guided by Shading Supervision

Previous SL-PSNs do not reveal the deep features and fail to explicitly leverage them to facilitate resolving the PS challenges. In contrast, based on insights from Section 3.2 that shading has a strong correlation with deep features,

7. The data flow of light-aware deep features is provided in the supplementary material, demonstrating that the output light-free deep features are unaffected by the order of lights.

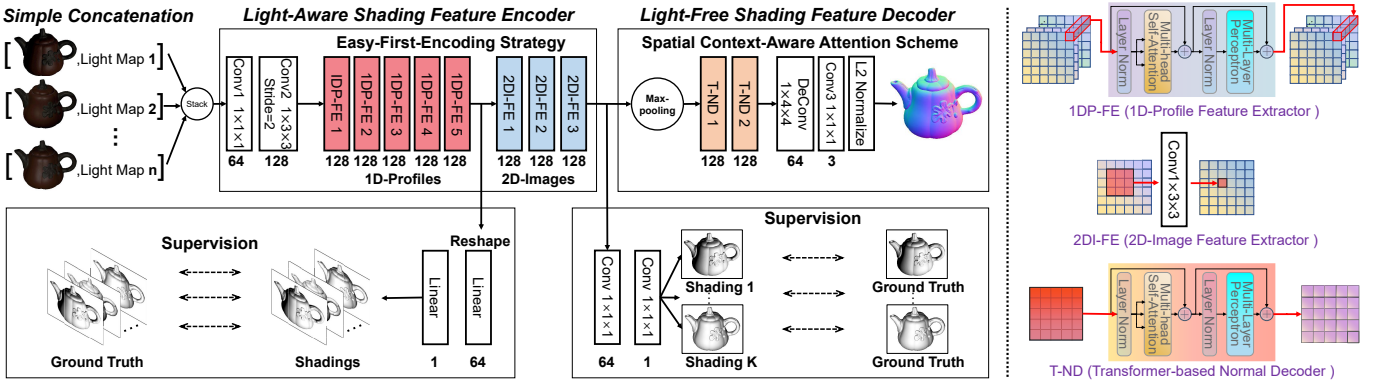


Fig. 8. The overall architecture of the proposed ESSENCE-Net (taking the per-first-encoding strategy as example). The values below each layer represent the number of the output feature channels.

we introduce shading supervision for 1D-profiles and 2D-images. An MLP is used to obtain accurate representations of shading, which have been proven to enhance feature quality by mapping features to the space of contrastive loss in contrastive learning [64], [65], [66]. Specifically, the deep shading features extracted from both are processed through an MLP to obtain shading, which are used to calculate shading loss. For shading features extracted from 1D-profiles with dimensions of $128 \times K \times H/2 \times W/2$, we first reshape them into $HW/4 \times K \times 128$, then apply two linear layers: the first reduces the dimensions to $HW/4 \times K \times 64$ and the second reduces it to $HW/4 \times K \times 1$. We then reshape the output to $K \times H/2 \times W/2$ for loss calculation with the ground truth shading (resized to $K \times H/2 \times W/2$). For shading features extracted from 2D-images with dimensions of $128 \times K \times H/2 \times W/2$, we apply two 3D convolutional layers to first reduce the dimensions to $64 \times K \times H/2 \times W/2$ and then to $K \times H/2 \times W/2$, allowing for loss calculation with the ground truth shading (resized to $K \times H/2 \times W/2$). We adopt the mean squared error to minimize the gap between the extracted and ground truth shadings, as follows

$$\mathcal{L}_{\text{1DP}} = \frac{1}{KHW} \sum_{k} \sum_{h,w} (S_{k,h,w} - \widehat{S}_{k,h,w})^2, \tag{4}$$

$$\mathcal{L}_{\text{2DI}} = \frac{1}{KHW} \sum_{k} \sum_{h,w} (S_{k,h,w} - \widehat{S}_{k,h,w})^2, \tag{5}$$

where $\mathcal{L}_{\text{1DP}}$ and $\mathcal{L}_{\text{2DI}}$ represent the shading loss for 1D-profiles and 2D-images, S and $\widehat{S}$ denote the extracted and ground truth shading.

Additionally, we apply the commonly used cosine similarity loss to minimize the error between the predicted and ground truth normal map, which is formulated as

$$\mathcal{L}_{\text{normal}} = \frac{1}{HW} \sum_{h,w} (1 - N_{h,w} \cdot \widehat{N}_{h,w}), \tag{6}$$

where N and $\widehat{N}$ denote the predicted and ground truth normal, and $\cdot$ indicates the dot product operation.

We empirically assign weights (0.25) to shading loss. The overall loss function is

$$\mathcal{L} = 0.25 \times \mathcal{L}_{\text{1DP}} + 0.25 \times \mathcal{L}_{\text{2DI}} + \mathcal{L}_{\text{normal}}. \tag{7}$$

By integrating shading loss, the network encodes more accurate shading features and achieves faster convergence.

4.2 Deep Feature Encoding by Easy-First-Encoding Strategy

Previous hybrid methods employ different and fixed strategies to address the challenges of PS, which fail to achieve optimal performance and lack generalizability. By contrast, based on the analysis in Sections 3.2 and 3.3 regarding the strategy and network architectures for addressing these two challenges, we adopt an *easy-first-encoding* strategy and implement it using appropriate architectures (*i.e.*, transformer for 1D-profiles and CNN for 2D-images).

First, we employ a 3D CNN layer with kernel size $1 \times 1 \times 1$ to increase dimensionality (from $6 \times K \times H \times W$ to $64 \times K \times H \times W$), followed by a downsampling layer to reduce feature size (from $64 \times K \times H \times W$ to $128 \times K \times H/2 \times W/2$). To encode shading features from 1D-profiles, we employ five transformer layers (each with a 4-head self-attention and an MLP) along the light (2nd) dimension, as analyzed in Section 3.3.1, detailed in Fig. 9. Transformer can maintain light permutation invariance compared to CNN [26] and can prevent rank collapse compared to self-attention [63], thus achieving better results. To encode shading features from 2D-images, we employ three 3D CNN layers with kernel size $1 \times 3 \times 3$, each followed by a Leaky ReLU layer, as analyzed in Section 3.3.1. Both the transformer for encoding features from 1D-profiles and the 3D CNN for encoding features from 2D-images do not alter the feature dimensions (*i.e.*, $128 \times K \times H/2 \times W/2$).

4.3 Deep Feature Decoding by Spatial Context-Aware Attention

Previous methods have extensively explored network architectures for encoding deep features but pay little attention to the network architectures for deep feature decoding, as analyzed in Section 3.3. Thus, we focus more on the network architectures of *light-free shading feature decoder*. Previous per-pixel methods lacked the utilization of spatial information in normal decoding. All-pixel and hybrid methods using CNN architectures could only leverage local shape similarities to decode normals due to the limited receptive fields [59]. In

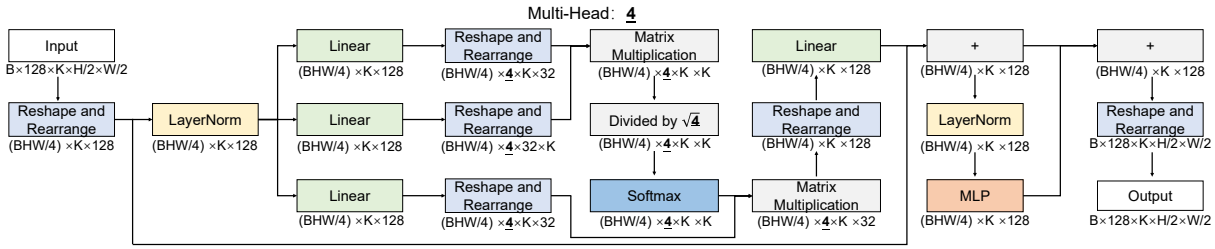


Fig. 9. Implementation details of transformer in encoding deep features from 1D-profiles (1DP-FE). The notation below each block represents the output feature size. B represents the batch size, 128 is the number of feature channels, H and W are the image width and height, and K is the number of lights. The number $\underline{4}$ indicates the number (4) of the heads in multi-head self-attention.

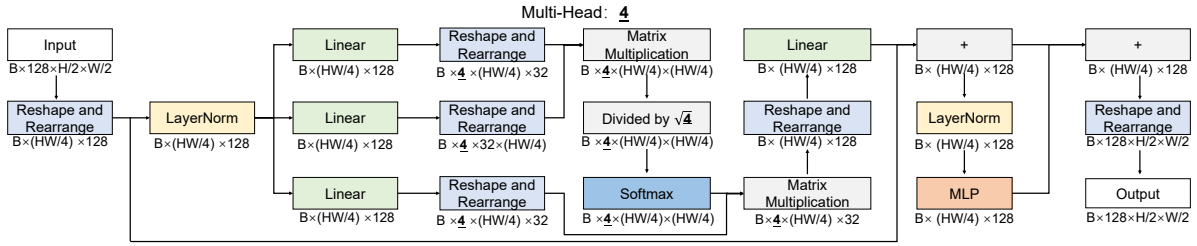


Fig. 10. Implementation details of transformer in decoding the normal map from deep shading features (T-ND). The notation below each block represents the output feature size. B represents the batch size, 128 is the number of feature channels, H and W are the image width and height. The number $\underline{4}$ indicates the number (4) of the heads in multi-head self-attention.

contrast, based on the analysis in Section 3.3.2 that spatial context-aware attention can leverage shape similarity globally, we employ transformer to decode the normal map from deep shading features. In *light-free shading feature decoder*, the deep shading features are first fed into a max-pooling layer to remove light information (from $128 \times K \times H/2 \times W/2$ to $128 \times H/2 \times W/2$). Subsequently, two transformer layers are employed to decode normal features, which utilize self-attention mechanisms to identify regions with similar shapes globally along the spatial dimensions (2nd and 3rd), as detailed in Fig. 10. Since the features have been downsampled on 2D-images ($1/4$) and 1D-profiles ($1/K$), it is feasible to utilize the spatial context-aware attention of transformer with lower computational cost. Finally, an upsampling layer followed by a 3D CNN layer with kernel size $1 \times 1 \times 1$ and L2 normalization is used to output an accurate normal map (from $128 \times H/2 \times W/2$ to $3 \times H \times W$).

5 Network Analysis and Ablation Study

In this section, we analyze the proposed ESSENCE-Net to validate the effectiveness of the proposed shading supervision, *easy-first-encoding* strategy, and network architectures. We retrain all the alternatives of ESSENCE-Net to ensure the validity of ablation studies by carefully controlling variations in each alternative and keeping other parts fixed.

Implementation Details. ESSENCE-Net contains $1,365,378$ parameters and is implemented using PyTorch with the Adam optimizer. It is trained for approximately 8 hours with an NVIDIA GeForce RTX 3090 GPU on synthetic datasets proposed in [20], following the training settings in [20]. The mean angular error (MAE) between estimated and ground truth normals is used as the evaluation metric.

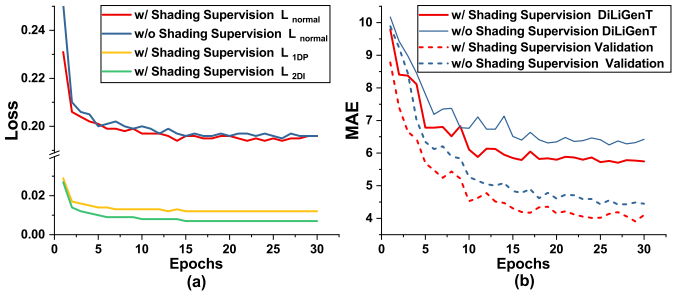


Fig. 11. The effect of shading supervision on convergence. (a) Manifestation of convergence in loss. (b) Manifestation of convergence in validation and *DiLiGenT* [46] datasets.

5.1 Validation of the Effectiveness on Shading Supervision for Deep Feature

5.1.1 Ablation Experiments

To validate the effectiveness of the proposed shading supervision strategy, we conducted ablation experiments on the *DiLiGenT* dataset [46]. The experiments with IDs 1 and 2 in Tab. 2 compare the performance of the proposed method *w/* and *w/o* shading supervision. Adding shading as an intermediate supervision improves the performance of nearly all objects. The second column in Fig. 15 shows the shading features extracted by the proposed ESSENCE-Net. The shading information is close to the ground-truth, whether inferred from 1D-profiles or 2D-images, which indicates the effectiveness of shading supervision.

5.1.2 Effectiveness on Convergence

To validate the effectiveness of the proposed shading supervision in facilitating convergence, we visualized the changes

TABLE 2
Quantitative Comparison of Results for Alternatives of ESSENCE-Net on the *DiLiGenT* Benchmark Dataset [46].

ID	Normalization	Shading Supervision	Strategy	Encoder (1D-Profiles)	Encoder (2D-Images)	Decoder	BALL	CAT	POT1	BEAR	POT2	BUDDHA	GOBLET	READING	COW	HARVEST	AVG.
1	Max	✓	Per-First-Encoding	Transformer	CNN	Transformer	2.11	4.23	4.94	4.26	4.90	6.25	6.32	7.72	4.99	11.22	5.69
2	Max	✗	Per-First-Encoding	Transformer	CNN	Transformer	2.40	4.25	5.98	4.82	6.52	6.44	6.88	8.64	5.14	11.43	6.25
3	Max	✓	All-First-Encoding	Transformer	CNN	Transformer	2.12	5.04	5.32	4.74	7.05	7.03	8.03	9.43	6.83	13.58	6.92
4	Max	✓	Per-All-Parallel	Transformer	CNN	Transformer	2.66	5.29	6.34	4.25	6.36	6.46	6.63	7.36	6.90	11.57	6.38
5	Max	✓	Per-All-Alternate	Transformer	CNN	Transformer	2.73	4.96	5.81	4.68	5.50	6.52	6.89	8.41	5.95	12.60	6.41
6	Max	✓	Per-First-Encoding	Transformer	CNN	CNN	2.60	4.77	6.35	4.23	5.88	6.67	6.80	8.48	5.43	11.72	6.29
7	Max	✓	Per-First-Encoding	Self-Attention	CNN	Transformer	2.14	4.90	5.81	4.25	5.93	6.67	7.43	8.02	5.17	12.11	6.24
8	Max	✓	Per-First-Encoding	CNN	CNN	Transformer	3.69	5.50	6.74	6.23	7.80	7.99	10.17	10.81	8.32	14.67	8.19
9	L2	✓	Per-First-Encoding	Transformer	CNN	Transformer	2.21	5.09	5.54	5.01	5.43	6.51	6.60	8.43	5.45	10.51	6.08
10	*w/o*	✓	Per-First-Encoding	Transformer	CNN	Transformer	2.95	5.66	6.62	4.01	5.94	7.69	8.29	15.51	5.10	14.48	7.62
11*	Max	✓	Per-First-Encoding	Transformer	CNN	Transformer	2.34	4.55	5.77	4.32	5.68	6.60	6.92	8.96	5.88	11.56	6.26
12*	Max	✓	Per-First-Encoding	Transformer	Transformer	Transformer	2.35	4.50	5.57	3.96	5.93	6.57	6.76	8.63	5.43	13.12	6.28

*The training size for experiments with IDs 11 & 12 is halved compared to other experiments to accommodate the usage of transformer in 2D-images.

TABLE 3
Performance of ESSENCE-Net *w/* and *w/o* Shading Supervision on $SynTest^{MERL}$ Dataset, Showing the Average MAE of 100 Materials.

Method	SPHERE	BUNNY	ARMADILLO	DRAGON
w/o Shading Supervision	2.71	2.80	3.64	3.55
w/ Shading Supervision	2.39	2.56	3.30	3.17

in loss and normal estimation performance during the training process. Fig. 11a illustrates the change in normal loss ($\mathcal{L}_{\text{normal}}$) and shading loss ($\mathcal{L}_{\text{1DP}}$, $\mathcal{L}_{\text{2DI}}$) during the training process *w/* and *w/o* shading supervision. Shading supervision significantly accelerates the convergence speed, and the shading loss of 1D-profiles ($\mathcal{L}_{\text{1DP}}$) and 2D-images ($\mathcal{L}_{\text{2DI}}$) is markedly decreasing and maintained at a low level. Fig. 11b shows the convergence on the validation and *DiLiGenT* [46] datasets. Shading supervision significantly speeds up the convergence and achieves better performance. These experimental results indicate that shading supervision enables ESSENCE-Net to not only learn more accurate shading features but also accelerate convergence speed.

5.1.3 Performance on Data with Different Materials

To evaluate the performance of the proposed shading supervision strategy across various materials, we conducted experiments on $SynTest^{MERL}$ dataset, which includes SPHERE, BUNNY, ARMADILLO, and DRAGON objects, each with 100 materials from the *MERL* dataset [67]. Tab. 3 shows an improvement of the proposed shading supervision by a comparison between *w/* and *w/o* shading supervision[8], validating the effectiveness of the shading supervision strategy.

5.2 Validation of the Effectiveness on Easy-First-Encoding Learning Strategy for Deep Feature Encoding

To validate the effectiveness of the proposed *easy-first-encoding* learning strategy, we compared it with other strategies across Case I to Case IV as classified in Section 3.2.

Validation on different resolutions: Case I to Case III. According to the analysis in Section 3.2, extracting shading features from 1D-profiles is consistently relatively easy from Case I to Case III. Figs. 12a and 12b show the performance of different strategies as the image resolution increases (32 and 96 images), with the models retrained using the corresponding image resolutions. The *easy-first-encoding* strategy (per-first-encoding) outperforms other strategies.

Validation on different numbers of images: Case II to Case III. According to the analysis in Section 3.2, from Case II to Case III, the ease of extracting shading features from 2D-images will be surpassed by that of 1D-profiles. We retrain the models and test their performance using the same number of images (less than 32). Fig. 12c shows that as the image number increases to a certain number (7), the performance of per-first-encoding gradually surpasses the all-first-encoding *w/o* normalization, indicating the validity of *easy-first-encoding* strategy. Moreover, *w/* normalization, this surpassing occurs earlier because normalization reduces the difficulty of extracting shading features from 1D-profiles. And experiments with IDs 1, 3, 4, and 5 in Tab. 2 demonstrate the superiority of the *easy-first-encoding* strategy under dense inputs. Fig. 12d further illustrates the performance of the proposed *easy-first-encoding* strategy on *DiLiGenT10*[2] dataset [47], where the per-first-encoding strategy surpasses the all-first-encoding strategy at 11 and 6 images, *w/o* and *w/* normalization, respectively. Fig. 13a further elucidates the specific performance of the *easy-first-encoding* strategy, showing that *w/o* normalization, the advantage of the per-first-encoding strategy becomes more apparent (blue) as the number of images increases, especially on isotropic and simple-shaped surfaces, as extracting shading features from 1D-profiles becomes easier with more images[9]. In contrast, normalization reduces the difficulty of extracting shading features from 1D-profiles and 2D-images, positively affecting the *easy-first-encoding* strategy.

Validation on SVBRDF surfaces: Case IV. According to the analysis in Section 3.2, no-uniform materials make feature extraction from 2D-images more challenging. Fig. 14 shows that the performance of *easy-first-encoding* strategy (per-first-encoding) far exceeds that of comparative methods all-first-encoding, per-all-alternate, and per-all-parallel. Fig. 15 shows that shading features extracted by per-first-encoding are significantly closer to ground truth. In contrast, shading features extracted in all-first-encoding and per-all-parallel orders are significantly affected by non-uniform materials.

Effectiveness of normalization strategy. According to the analysis in Section 3.2, normalization preprocessing significantly reduces the difficulty of extracting shading features, especially for isotropic and simple-shaped surfaces. Fig. 12 demonstrates that normalization consistently improves the performance of normal estimation, regardless of the strategy or case. Fig. 13b illustrates the impact of normalization on per-first-encoding and all-first-encoding strategies across

8. For more detailed performance on the 100 materials, please refer to the supplementary material.

9. For more detailed performance of the *easy-first-encoding* strategy on *DiLiGenT10*[2] dataset [47], please refer to the supplementary material.

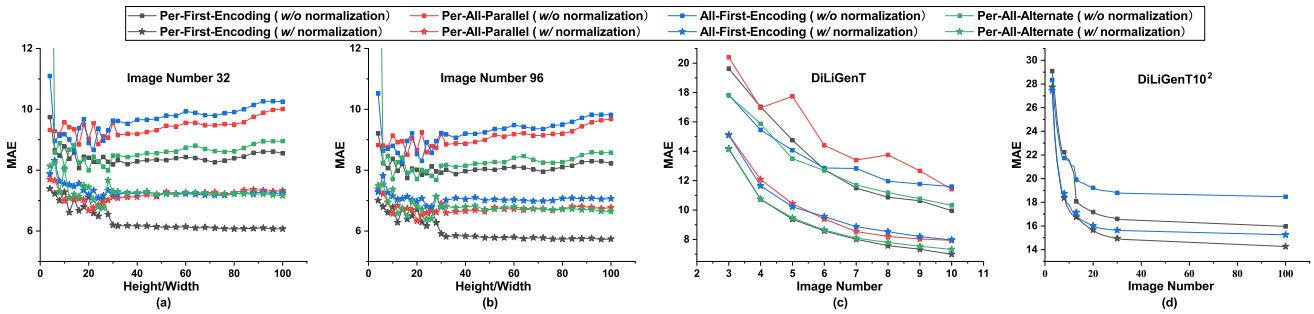


Fig. 12. Validation of the *easy-first-encoding* strategy across different resolutions and numbers of images. (a) and (b) Comparison of the performance of different learning strategies at different resolutions with 32 and 96 images on the *DiLiGenT* dataset [46]. (c) and (d) Comparison of the performance of different learning strategies with varying numbers of inputs on the *DiLiGenT* [46] and *DiLiGenT10²* [47] datasets.

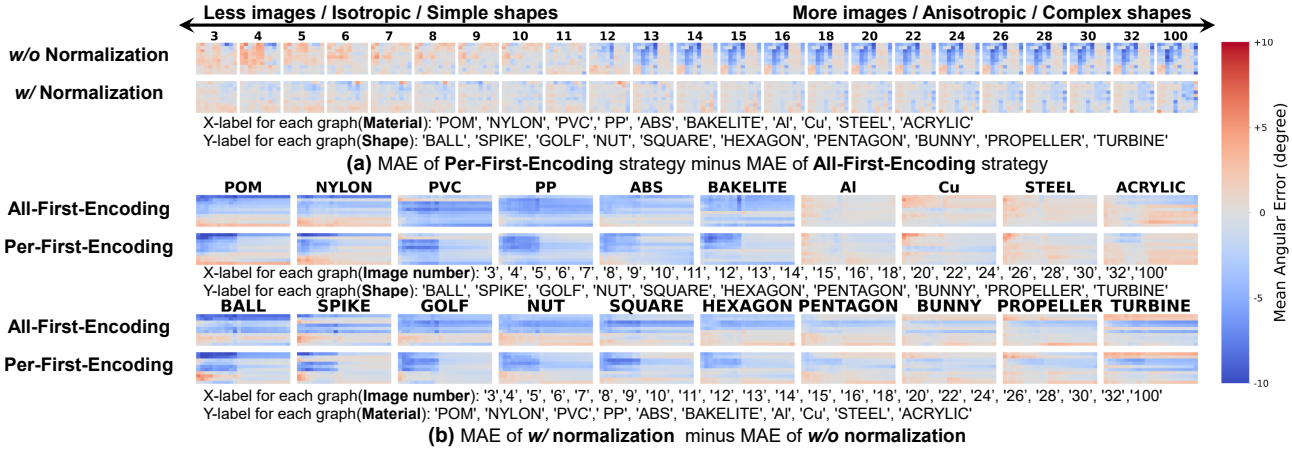


Fig. 13. The specific performance of the *easy-first-encoding* strategy and normalization strategy on the *DiLiGenT10²* dataset [47]. Each graph in the figure represents the performance on a specific image number, shape, or material, with the x-axis as the horizontal axis and the y-axis as the vertical axis. (a) Error matrix of the per-first-encoding strategy's MAE minus all-first-encoding strategy's MAE. (b) Error matrix of the *w/* normalization strategy's MAE minus *w/o* normalization strategy's MAE.

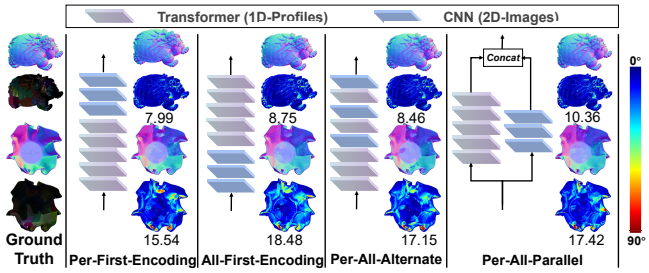


Fig. 14. Performance of different learning strategies on SVBRDF surfaces from the *CyclesPS* dataset [14].

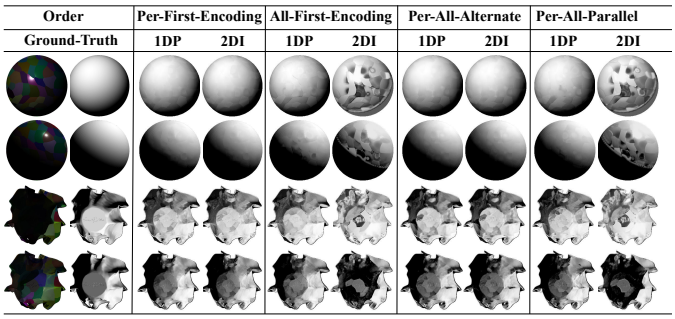


Fig. 15. Shading features extracted from 1D-profiles (1DP) and 2D-images (2DI) on *CyclesPS* [14] for different strategies.

various surfaces, showing significant effects on isotropic and simple-shaped surfaces while still beneficial but less prominent for anisotropic and complex-shaped surfaces[10]. Experiments in Tab. 2 with IDs 1, 9, and 10 compare the effects of commonly used max normalization, L2 normalization, and *w/o* normalization, indicating that max normalization better facilitates the extraction of shading features.

10. For more detailed performance of the normalization strategy on *DiLiGenT10²* dataset [47], please refer to the supplementary material.

5.3 Validation of the Effectiveness on Network Architectures

5.3.1 *Effectiveness of Transformer in 1D-Profiles*

To validate the effectiveness of the transformer architecture in 1D-profiles, we compared its performance with CNN and self-attention on the *DiLiGenT* dataset [46]. We also conducted comparative analyses of the attention maps between transformer and self-attention. Experiments with IDs 1, 7, and 8 in Tab. 2 show that transformer outperforms CNN

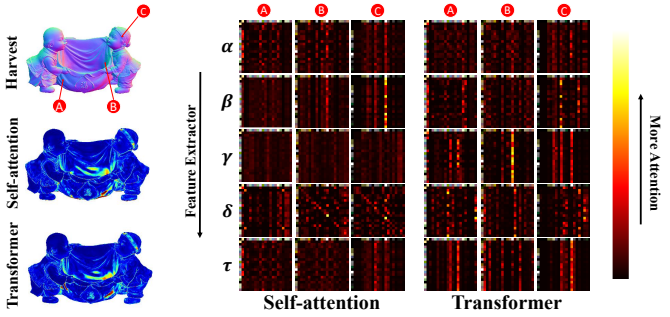


Fig. 16. Attention maps of transformer and self-attention on extracting features from 1D-profiles. α, β, γ ,δ, and τ are five consecutive transformer or self-attention layers. The first row and column of each attention map represent the values of the 1D-profiles.

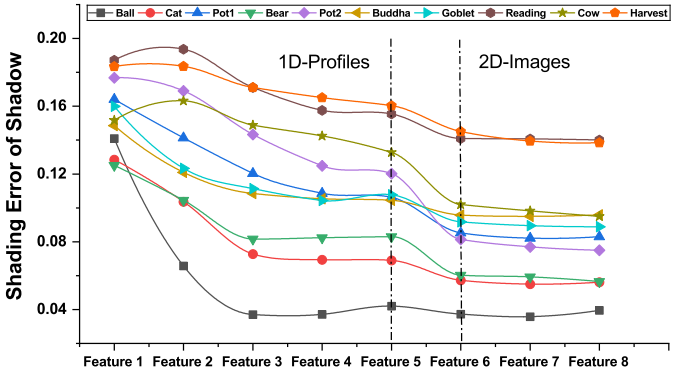


Fig. 17. The shading error variation of the proposed ESSENCE-Net in shadow regions as the network goes deeper.

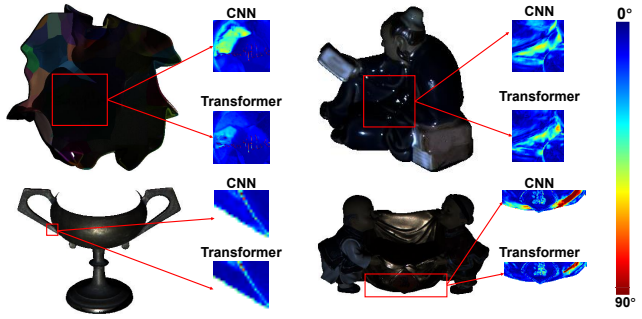


Fig. 18. Comparison between the transformer and CNN for normal decoding in regions with severe shadows.

and self-attention almost on all objects. Fig. 16 visualizes the attention and error maps for self-attention and transformer on HARVEST with 20 images. It shows that self-attention assigns high attention values on the diagonal line of the attention maps, *e.g.*, δ layer, while transformer does not, suggesting that self-attention focuses less on global information exchange within 1D-profiles than transformer. Additionally, self-attention has greater attention on shadow regions, *e.g.*, point C in the attention maps of the γ and τ layers, while transformer relies more on information from general reflection points to encode shading features. The error map also indicates that transformer better handles general reflectance and shadow areas. These experimental results indicate that transformer can more effectively leverage global information from 1D-profiles to address unknown reflectance issues for encoding shading features.

5.3.2 Effectiveness of CNN in 2D-Images

To validate the effectiveness of CNN architecture in 2D-images, we evaluate its performance in handling shadows and compare its normal estimation performance with that of transformer on the *DiLiGenT* dataset [46]. We compute the shading errors in shadow regions (threshold set at 10^{-6} [40]) for each layer, as shown in Fig. 17. As the processing shifts from 1D-profiles to 2D-images, there is a significant decrease in shading error, validating the hypothesis in Section 3.3.1 that global illumination effects can be complemented by shape clues and the cross-correlation principle of CNN. Experiments with IDs 11 and 12 in Tab. 2 compares the performance of both approaches when training sizes are halved. The performance of CNN is not inferior to that of transformer, validating its effectiveness in encoding shading features from 2D-images.

5.3.3 Effectiveness of Spatial Context-Aware Attention Scheme in Shading Feature Decoding

To validate the effectiveness of the proposed spatial context-aware attention in normal decoding, we compare its performance with CNN on the *DiLiGenT* dataset [46] and visualized the attention maps of objects from *CyclesPS* [14] for in-depth analysis. Experiments with IDs 1 & 6 in Tab. 2 compare the performance of different decoding methods. With spatial context-aware attention, transformer performs better on most objects than CNN, whether on objects with general reflectance (*e.g.*, BALL, POT1) or objects with significant global illumination effects (*e.g.*, READING). We select some non-convex objects of *DiLiGenT* [46] and *CyclesPS* [14] datasets to show the performance in shadow areas in Fig. 18. Using transformer with spatial context-aware attention for normal decoding results in significantly better performance in shadow regions. Fig. 19 shows the attention maps of transformer when decoding surface normals. Transformer focuses not only on the neighborhood information of the target point but also on regions with similar normals. Moreover, for regions susceptible to shadow effects, such as point C on PIG, there is a lack of effective information within the neighborhood. Transformer primarily seeks points with similar shapes globally; for instance, there is a greater focus on the lower left part of the face, aiding in decoding normal for point C in the ear region. Consequently, the proposed spatial context-aware attention can decode normal maps more accurately by leveraging shape similarity globally.

5.3.4 Effects of the Network Architecture Number

To determine the proper block numbers of 1DP-FE (1D-profile feature extractor), 2DI-FE (2D-image feature extractor), and T-ND (transformer-based normal decoder), we present both training metrics (training time, resource consumption, and MAE on the validation dataset during training) and testing performance (MAE on *DiLiGenT* [46], *DiLiGenT10*^{2} [47], and *DiLiGenT-II* [48] datasets) in Fig. 20.

Number of 1DP-FE: As shown in Fig. 20a, increasing the number of 1DP-FE does not result in continuous improve-

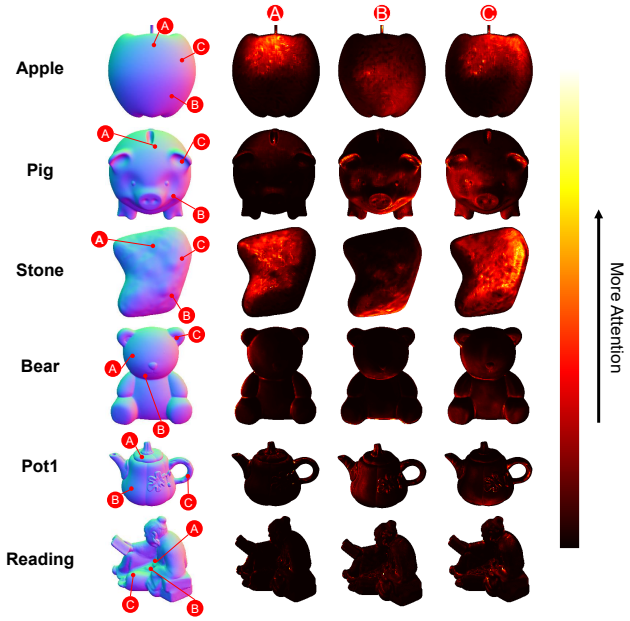


Fig. 19. Visualization of the attention maps for the transformer decoder.

ments on the validation dataset, while both training time and memory consumption increase. We therefore choose 5 1DP-FE blocks for encoding deep features from 1D-profiles. Furthermore, as shown in Fig. 20b, the MAE on the three benchmark datasets does not consistently decrease with more 1DP-FE blocks, validating that 5 1DP-FE blocks are enough for deep feature encoding and that additional blocks may lead to overfitting.

Number of 2DI-FE: As shown in Fig. 20c, during training, the MAE on the validation dataset significantly decreases when the number of 2DI-FE increases to 3, after which it stabilizes, while the training time and memory consumption continue to rise. Thus, we select 3 2DI-FE blocks to extract deep features from 2D-images. As shown in Fig. 20d, the MAE on the benchmark datasets further validates the optimal performance with 3 2DI-FE blocks.

Number of T-ND: As shown in Fig. 20e, 2 T-ND blocks already show excellent performance on the validation dataset. Thus, we choose 2 T-ND blocks for normal decoding which avoids excessive computational resources at the same time. The performance on benchmark datasets in Fig. 20f confirms that 2 T-ND blocks achieve optimal performance on the benchmark datasets.

6 Overall Performance Experiments

This section presents the overall performance of the proposed ESSENCE-Net on the pre-trained models and compares it with state-of-the-art methods on three real-world datasets. We train ESSENCE-Net with 32 input images for dense setups and 10 input images for sparse setups[11]. We evaluate ESSENCE-Net on real-world datasets *DiLiGenT* [46], *DiLiGenT*10^2 [47], and *DiLiGenT-II* [48].

11. The differing performances between models trained with 10 images and 32 images can be found in the supplementary materials.

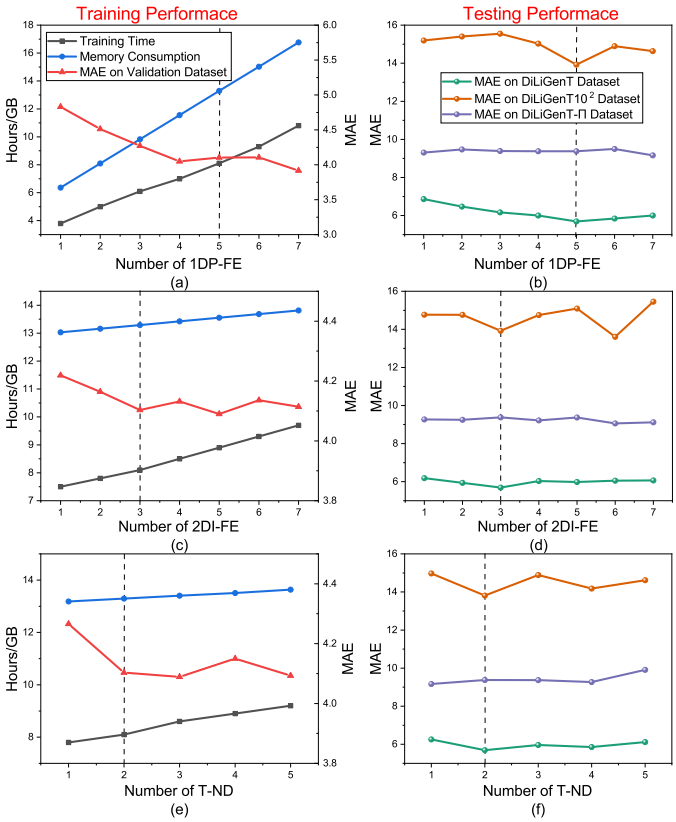


Fig. 20. The training performance (training time, resource consumption, and MAE on the validation dataset) and testing performance (MAE on *DiLiGenT* [46], *DiLiGenT10*^{2} [47], and *DiLiGenT-II* [48] datasets) under different numbers of 1DP-FE, 2DI-FE, and T-ND blocks.

6.1 Performance on *DiLiGenT* Dataset

DiLiGenT dataset [46] consists of 10 objects with 96 images, commonly used for evaluating photometric stereo methods.

6.1.1 Dense Input

To validate the effectiveness of the proposed ESSENCE-Net in addressing unknown reflectance and global illumination effects, we compare its performance against conventional methods (L2 [1], IA14 [68], ST14 [40]), per-pixel SL-PSNs (DPSN [69], CNN-PS [14], PX-Net [17]), all-pixel SL-PSNs (PS-FCN^{+N} [20], NormAttention-PSN [24], Wang20 [30]), and hybrid SL-PSNs (GPS-Net [28], SPS-Net [26], MT-PS-CNN [25], HT21 [70], Uni MS-PS [71]) on the *DiLiGenT* [46] dataset with 96 input images[12], as shown in Tab. 4. Compared to previous methods, ESSENCE-Net shows superior performance on general reflective surfaces (*e.g.*, CAT and COW) and complex surfaces heavily influenced by global illumination effects (*e.g.*, READING and HARVEST). Benefiting from the proposed shading features encoding and decoding framework, ESSENCE-Net achieves state-of-the-art performance with an average MAE of 5.69°. The experimental results validate the effectiveness of ESSENCE-Net in handling the challenges posed by unknown reflectance and global illumination effects.

12. For qualitative comparison on the *DiLiGenT* [46] dataset with dense inputs, please refer to the supplemental material.

TABLE 4
Comparison of the Proposed ESSENCE-Net with Existing State-of-the-art Methods in a Dense Setting on the *DiLiGenT* Benchmark Dataset [46].

Category	Method	BALL	CAT	POT1	BEAR	POT2	BUDDHA	GOBLET	READING	COW	HARVEST	AVG.
Conventional	L2 [1]	4.10	8.41	8.89	8.39	14.65	14.92	18.50	19.80	25.60	30.62	15.39
	IA14 [68]	3.34	6.74	6.64	7.11	8.77	10.47	9.71	14.19	13.05	25.95	10.60
	ST14 [40]	1.74	6.12	6.51	6.12	8.78	10.60	10.09	13.63	13.93	25.44	10.30
Per-pixel	DPSN [69]	2.02	6.54	7.05	6.31	7.86	12.68	11.28	15.51	8.01	16.86	9.41
	CNN-PS [14]	2.12	4.38	5.37	4.20	6.38	8.07	7.42	12.12	7.92	14.08	7.21
	PX-Net [17]	1.95	4.26	4.86	3.46	5.03	7.58	6.71	9.78	4.72	13.30	6.17
All-pixel	PS-FCN^{+N} [20]	2.67	4.73	6.15	5.01	7.15	7.56	7.88	10.98	6.70	12.42	7.12
	NormAttention-PSN [24]	2.93	4.65	5.96	4.80	6.42	7.12	7.49	9.93	5.99	12.28	6.76
	Wang20 [30]	1.78	4.66	6.46	4.12	6.45	6.09	7.22	10.05	6.33	13.34	6.65
Hybrid	GPS-Net [28]	2.92	5.42	6.04	5.07	7.01	7.77	9.00	13.58	6.14	15.14	7.81
	SPS-Net [26]	2.80	5.10	7.50	-	7.40	6.90	7.10	11.90	6.30	13.70	7.60
	MT-PS-CNN [25]	2.29	5.87	6.92	5.79	6.89	6.85	7.88	11.94	7.48	13.71	7.56
	HT21 [70]	2.30	4.50	5.03	3.66	5.03	7.22	6.95	11.30	4.85	12.70	6.35
	Uni MS-PS§ [71]	1.93	3.05	3.85	2.64	4.32	5.88	6.40	7.31	3.76	10.44	4.96
	ESSENCE-Net	2.11	4.23	4.94	4.26	4.90	6.25	6.32	7.72	4.99	11.22	5.69

Results with § are obtained with 30 images as input. The first 20 images of BEAR are discarded in CNN-PS [14], PX-Net [17], PS-FCN^{+N} [20], NormAttention-PSN [24], Wang20 [30], GPS-Net [28], HT21 [70], and ESSENCE-Net.

TABLE 5
A Comparison of the Proposed ESSENCE-Net and Uni MS-PS [71] in Terms of Training Parameters and Testing Time on the *DiLiGenT* Benchmark Dataset [46] (with 96 Input Images).

Method	Training			Testing Time (second)†										
	Parameters	Materials	Time (hour)	BALL	CAT	POT1	BEAR	POT2	BUDDHA	GOBLET	READING	COW	HARVEST	Average
Uni MS-PS [71]	80M	200000	72‡	73.43	593.79	591.52	74.08	592.94	588.09	583.58	74.76	72.73	587.83	383.28
ESSENCE-Net	1M	100	8†	0.38	1.26	1.46	0.74	1.03	1.01	1.31	0.74	0.65	1.28	0.99

Results with † are obtained on an NVIDIA GeForce RTX 3090 GPU (24GB), while results with ‡ are obtained on an NVIDIA A100 GPU (80GB).

As shown in Tabs. 4 and 5, the proposed ESSENCE-Net demonstrates competitive performance compared to Uni MS-PS [71], despite being trained with significantly fewer materials (100 materials for ESSENCE-Net vs. 200,000 materials for Uni MS-PS [71], dataset not released). This highlights the effectiveness of the proposed strategies, including shading supervision, *easy-first-encoding* for deep shading feature encoding, and spatial context-aware attention for normal decoding, in achieving optimal normal estimation accuracy. Moreover, the proposed ESSENCE-Net excels with its lightweight design, featuring only 1 million learnable parameters, which is two orders of magnitude fewer than Uni MS-PS (80M) [71]. This leads to significantly faster training speed (8 RTX3090 hours for ours vs. 72 A100 hours for Uni MS-PS [71]) and significantly faster testing speed (0.99 seconds for ours vs. 383.28 seconds for Uni MS-PS [71]). The remarkable training efficiency and high accuracy of ESSENCE-Net make it highly competitive for real-world applications, such as industrial quality inspection [2], where efficiency is of utmost importance.

6.1.2 Sparse Input

Under a sparse set of lights, the limited information in 1D-profiles makes normal estimation challenging. Tab. 6 compares the proposed ESSENCE-Net with state-of-the-art sparse photometric stereo methods [14], [16], [17], [18], [20], [26], [27], [28], [30] on the *DiLiGenT* [46] dataset with 10 input images[13]. We randomly sample 10 out of 96 images for normal estimation. We repeat this process 100 times and report the average MAE. ESSENCE-Net outperforms all methods with an average MAE of 6.98°, even surpassing most methods under dense inputs (Tab. 4), indicating that ESSENCE-Net can effectively handle unknown reflectance and global illumination effects even under sparse setups.

13. For qualitative comparison on the *DiLiGenT* [46] dataset with sparse inputs, please refer to the supplemental material.

6.2 Performance on *DiLiGenT10^2* Dataset

To evaluate the performance of the proposed ESSENCE-Net across various material and shapes, we conducted experiments on the *DiLiGenT10^2* dataset [47], which comprises 100 objects with 10 shapes and 10 materials. The comparison between ESSENCE-Net and existing state-of-the-art methods is illustrated in Fig. 21. Benchmark results of the comparative methods (L2 [1], ST14 [40], PS-FCN [19], CNN-PS [14], SPLINE-Net [18], GPS-Net [28], and PX-Net [17], Uni MS-PS [71]) are collected from [47] or the authors. The proposed ESSENCE-Net performs remarkably well across various reflective and geometric surfaces, achieving an average MAE of 13.81°, indicating its effectiveness in handling various unknown reflectance and global illumination effects. Although the proposed ESSENCE-Net demonstrates slightly lower accuracy in normal estimation compared to Uni MS-PS [71] on the *DiLiGenT10^2* dataset [47], it significantly outperforms Uni MS-PS [71] in terms of efficiency, as described in Tab. 5.

6.3 Performance on *DiLiGenT-Π* Dataset

To evaluate the ability to estimate shape details of ESSENCE-Net, we conducted experiments on the *DiLiGenT-Π* dataset [48], which comprises 30 near-planar scenes with rich details featuring metallic, specular, rough, and translucent materials. The comparison between ESSENCE-Net and state-of-the-art methods is shown in Tab. 7. Benchmark results of comparative methods (L2 [1], WG10 [10], ST14 [40], PS-FCN [19], CNN-PS [14], PX-Net [17], and GPS-Net [28], Uni MS-PS [71]) are collected from [48] or the authors. ESSENCE-Net performs excellently across various

TABLE 6
Comparison of the Proposed ESSENCE-Net with Existing State-of-the-art Methods in a Sparse Setting (10 Input Images) on the *DiLiGenT* [46] Benchmark Dataset. All Results Are the Averages of 100 Random Experiments.

Method	BALL	CAT	POT1	BEAR	POT2	BUDDHA	GOBLET	READING	COW	HARVEST	AVG.
SPLINE-Net [18]	4.96	7.52	8.77	5.99	11.79	10.07	10.43	16.13	8.80	19.05	10.35
LMPS [16]	3.97	6.69	7.30	8.73	9.74	11.36	10.46	14.37	10.19	17.33	10.01
PS-FCN^{+N} [20]	4.26	6.43	7.84	6.52	9.96	9.52	10.32	12.86	12.06	17.50	9.72
SPS-Net [26]	4.60	6.90	8.90	–	9.00	8.00	9.00	13.60	8.30	16.70	9.44
GPS-Net [28]	4.33	6.81	7.50	6.34	8.38	8.87	10.79	15.00	9.34	16.92	9.43
CNN-PS [14]	4.06	6.54	6.94	5.79	8.23	9.97	10.45	13.63	9.88	17.44	9.29
PX-Net [28]	2.50	6.30	7.00	4.90	7.70	9.40	9.70	13.10	7.20	16.10	8.37
Wang20 [30]	2.30	5.62	7.08	5.18	8.19	7.05	8.80	10.88	7.53	15.26	7.79
PS-Transformer [27]	3.27	5.34	6.06	4.88	6.97	8.65	9.28	11.24	6.54	14.41	7.66
ESSENCE-Net	2.14	5.14	6.11	5.29	6.87	7.16	8.89	8.12	6.62	13.42	6.98

The first 20 images of BEAR are discarded, except for LMPS [16].

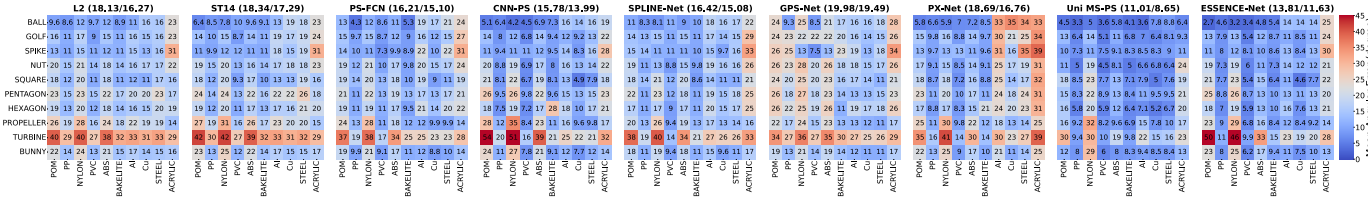


Fig. 21. Comparison of shape-material error matrix (mean/median) between ESSENCE-Net and state-of-the-art methods on *DiLiGenT10*^{2} [47].

TABLE 7
Comparison of the Proposed ESSENCE-Net with Existing State-of-the-art Methods on the *DiLiGenT-Π* Benchmark Dataset [48].

Method	Metallic										Specular										Translucent					Rough					AVG.
	FLOWER	BIRD	RHINO	LIONS	QUEEN	CRAB	SHIP	PARE	SAIL	FISH	TREE	OCEAN	LUNG	BEAR	TV	SUN	TCICHI	WAVE	ASTRO	WHALE	BAGUA-T	LOTUS-T	LION-T	PANDA-T	CLOUD-T	BAGUA-R	LOTUS-R	LION-R	PANDA-R	CLOUD-R	
L2 [1]	6.8	8.1	7.5	6.7	7.5	7.1	7.1	6.0	8.7	6.7	9.4	6.2	8.2	9.1	11.7	7.9	9.2	6.9	7.9	8.8	16.7	14.0	21.0	16.4	17.0	13.2	12.0	19.4	14.6	14.0	10.5
WG10 [10]	7.2	11.6	9.3	9.1	8.6	8.5	9.3	8.4	8.1	8.2	14.9	8.4	11.8	8.6	17.1	10.5	17.9	9.2	11.4	13.4	17.8	14.3	21.6	16.8	18.1	16.5	12.3	19.4	16.0	17.4	12.7
ST14 [40]	7.9	10.8	8.9	8.1	8.3	8.4	9.1	8.0	8.2	8.1	10.6	7.6	9.4	9.7	13.5	8.2	10.7	8.4	9.2	10.1	17.4	14.5	21.3	16.8	17.7	13.6	12.4	19.1	15.0	14.7	11.5
PX-Net [17]	5.5	8.8	10.6	6.8	14.7	7.2	12.3	4.9	13.4	7.4	11.3	5.9	7.6	7.9	12.1	6.7	8.7	7.1	7.1	9.5	15.8	13.7	18.4	15.9	16.1	11.9	11.0	17.9	14.1	13.6	10.8
CNN-PS [14]	4.7	6.8	4.9	4.7	5.4	4.5	4.9	3.9	5.2	4.2	7.8	4.6	5.7	7.4	11.3	5.8	8.3	5.3	6.0	11.6	16.4	13.5	20.3	16.6	17.2	12.2	10.9	15.8	14.2	14.6	9.2
PS-FCN [19]	4.6	7.2	5.3	4.5	4.7	5.3	6.1	4.7	5.1	4.6	10.2	5.8	9.7	7.4	10.6	6.7	8.0	6.8	7.2	12.2	16.8	13.6	21.2	17.2	17.8	13.0	11.8	18.4	14.8	14.3	9.8
GPS-Net [28]	4.6	6.8	5.6	4.6	4.7	4.9	5.1	4.0	5.5	4.6	9.3	5.4	7.5	6.9	10.6	8.0	8.3	6.3	7.7	10.4	16.1	13.7	23.4	17.0	17.6	16.4	13.4	23.0	16.3	15.9	10.1
Uni MS-PS§ [71]	4.9	6.6	5.2	5.4	5.7	5.8	6.3	4.7	6.7	4.1	6.1	5.4	5.4	6.7	10.2	5.9	7.3	5.5	6.0	6.7	11.8	8.1	11.4	11.4	11.0	9.6	10.2	13.1	13.0	12.6	7.8
ESSENCE-Net	4.4	6.8	5.6	5.3	6.5	4.3	4.7	4.1	5.5	4.5	10.3	5.4	7.3	7.7	9.9	7.4	9.3	6.1	7.0	10.1	16.6	13.5	20.2	16.7	17.4	11.8	10.6	15.6	13.8	13.3	9.4

Results with § are obtained with 30 images as input.

materials, achieving an average MAE of 9.4°. Despite being slightly lower than CNN-PS [14] (average MAE of 9.2°), ESSENCE-Net still achieves comparable performance. We hypothesize that this may be due to the spatial downsampling and upsampling operations, leading to a loss of details[14]. However, ESSENCE-Net demonstrates more efficient normal estimation than per-pixel CNN-PS [14], as analyzed in Section 2.1, and its performance remains equally outstanding. Additionally, although the proposed ESSENCE-Net demonstrates slightly lower accuracy in normal estimation compared to Uni MS-PS [71] on the *DiLiGenT-Π* dataset [48], it significantly outperforms Uni MS-PS [71] in terms of efficiency, as described in Tab. 5.

7 CONCLUSION

In this paper, we revisited the essence of SL-PSNs. We show that the SL-PSNs inherently learn features highly correlated to shading for normal recovery, *i.e.*, eliminating unknown reflectance and global illumination effects. We propose the shading supervision strategy and experimentally demonstrate its effectiveness in enhancing shading features and accelerating convergence. We propose an *easy-first-encoding* learning strategy based on analyzing the difficulty level in resolving unknown reflectance and global illumination effects across various cases, and experimentally demonstrate its ability to facilitate deep shading features extraction. We present an analysis of suitable network architectures for resolving unknown reflectance and global illumination effects to encode deep features and further propose spatial context-aware attention to decode the normal map from deep features. The experiments demonstrate that these architectures facilitate encoding deep features and accurately decode normals from them. We propose ESSENCE-Net based on shading supervision, *easy-first-encoding* deep feature encoding, and spatial context-aware attention-based normal decoding strategy. Experimental results have shown that the proposed method outperforms previous state-of-the-art methods.

Limitations. Although our method performs excellently in normal estimation, it has several limitations. First, ESSENCE-Net struggles with handling translucent materials (*e.g.*, ACRYLIC) and extremely complex shapes (*e.g.*, TURBINE), as shown in Fig. 21. Second, our method assumes directional lights and a camera with orthographic projection. Future work will focus on relaxing that assumption.

14. For a detailed illustration, please refer to the supplementary material.

ACKNOWLEDGMENTS

This work was supported by the Science and Technology Innovation Leading Talent Project of Special Support Plan for High-level Talents of Zhejiang Province under Grant No. 2022R52053, National Natural Science Foundation of China under Grant No. 52175520, 62376247, 61925603, 62136001, and 62088102, the Ministry of Education Singapore Tier 1 grant No. RG98/24, and National Key Research and Development Program of China under Grant No. 2023YFB3307202.

REFERENCES

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

[2] M. Ren, X. Wang, G. Xiao, M. Chen, and L. Fu, "Fast defect inspection based on data-driven photometric stereo," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 4, pp. 1148–1156, 2018.

[3] D. Miyazaki, K. Hara, and K. Ikeuchi, "Median photometric stereo as applied to the segonko tumulus and museum objects," *Int. J. Comput. Vis.*, vol. 86, pp. 229–242, 2010.

[4] B. Wu, Y. Li, W. C. Liu, Y. Wang, F. Li, Y. Zhao, and H. Zhang, "Centimeter-resolution topographic modeling and fine-scale analysis of craters and rocks at the chang'e-4 landing site," *Earth Planet. Sci. Lett.*, vol. 553, p. 116666, 2021.

[5] Georghiades, "Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 816–823.

[6] R. Ruiters and R. Klein, "Heightfield and spatially varying BRDF reconstruction for materials with interreflections," in *Comput. Graph. Forum*, 2009, pp. 513–522.

[7] S. Tozza, R. Mecca, M. Duocastella, and A. Del Bue, "Direct differential photometric stereo shape recovery of diffuse and specular surfaces," *J. Math. Imaging Vis.*, vol. 56, pp. 57–76, 2016.

[8] Q. Zheng, A. Kumar, B. Shi, and G. Pan, "Numerical reflectance compensation for non-lambertian photometric stereo," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3177–3191, 2019.

[9] Y. Mukaigawa, Y. Ishii, and T. Shakunaga, "Analysis of photometric factors based on photometric linearization," *JOSA A*, vol. 24, no. 10, pp. 3326–3334, 2007.

[10] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Proc. Asian Conf. Comput. Vis.*, 2011, pp. 703–717.

[11] T.-P. Wu and C.-K. Tang, "Photometric stereo via expectation maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 546–560, 2009.

[12] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa, "Robust photometric stereo using sparse regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 318–325.

[13] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep photometric stereo networks for determining surface normal and reflectances," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 114–128, 2020.

[14] S. Ikehata, "CNN-PS: CNN-based photometric stereo for general non-convex surfaces," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.

[15] Ikehata, Satoshi, "Does physical interpretability of observation map improve photometric stereo networks?" in *ICIP*, 2022, pp. 291–295.

[16] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita, "Learning to minify photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7568–7576.

[17] F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla, "PX-NET: Simple and efficient pixel-wise training of photometric stereo networks," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 12 757–12 766.

[18] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L.-Y. Duan, and A. C. Kot, "SPLINE-Net: Sparse photometric stereo through lighting interpolation and normal estimation networks," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8549–8558.

[19] G. Chen, K. Han, and K.-Y. K. Wong, "PS-FCN: A flexible learning framework for photometric stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.

[20] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Deep photometric stereo for non-lambertian surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 129–142, 2020.

[21] Y. Ju, K.-M. Lam, Y. Chen, L. Qi, and J. Dong, "Pay attention to devils: A photometric stereo network for better details," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2021, pp. 694–700.

[22] Y. Ju, J. Dong, and S. Chen, "Recovering surface normal and arbitrary images: A dual regression network for photometric stereo," *IEEE Trans. Image Process.*, vol. 30, pp. 3676–3690, 2021.

[23] Y. Ju, M. Jian, C. Wang, C. Zhang, J. Dong, and K.-M. Lam, "Estimating high-resolution surface normals via low-resolution photometric stereo images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2512–2524, 2024.

[24] Y. Ju, B. Shi, M. Jian, L. Qi, J. Dong, and K.-M. Lam, "NormAttention-PSN: A high-frequency region enhanced photometric stereo network with normalized attention," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 3014–3034, 2022.

[25] Y. Cao, B. Ding, Z. He, J. Yang, J. Chen, Y. Cao, and X. Li, "Learning inter-and intraframe representations for non-lambertian photometric stereo," *Opt. Lasers Eng.*, vol. 150, p. 106838, 2022.

[26] H. Liu, Y. Yan, K. Song, and H. Yu, "SPS-Net: Self-attention photometric stereo network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2020.

[27] Ikehata, Satoshi, "PS-Transformer: Learning sparse photometric stereo network using self-attention mechanism," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 1–12.

[28] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi, "GPS-Net: Graph-based photometric stereo network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 10 306–10 316.

[29] S. Ikehata, "Scalable, detailed and mask-free universal photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13 198–13 207.

[30] X. Wang, Z. Jian, and M. Ren, "Non-lambertian photometric stereo network based on inverse reflectance model with collocated light," *IEEE Trans. Image Process.*, vol. 29, pp. 6032–6042, 2020.

[31] S. Ikehata, "Universal photometric stereo network using global lighting contexts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12 591–12 600.

[32] Z. Li, Z. Lu, H. Yan, B. Shi, G. Pan, Q. Zheng, and X. Jiang, "Spin-UP: Spin light for natural light uncalibrated photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 11 905–11 914.

[33] Y. Hold-Geoffroy, P. Gotardo, and J.-F. Lalonde, "Single day outdoor photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2062–2074, 2019.

[34] S. Choi, S. Yoon, G. Nam, S. Lee, and S.-H. Baek, "Differentiable display photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 11 831–11 840.

[35] B. Yu, J. Ren, J. Han, F. Wang, J. Liang, and B. Shi, "EventPS: Real-time photometric stereo using an event camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[36] J. H. Chan, B. Yu, H. Guo, J. Ren, Z. Lu, and B. Shi, "ReLeaPS: Reinforcement learning-based illumination planning for generalized photometric stereo," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 9167–9175.

[37] A. Tiwari and S. Raman, "DeepPS2: Revisiting photometric stereo using two differently illuminated images," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 129–145.

[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[39] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv:2001.08361*, 2020.

[40] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1078–1091, 2013.

[41] Q. Zheng, B. Shi, and G. Pan, "Summary study of data-driven photometric stereo methods," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 3, pp. 213–221, 2020.

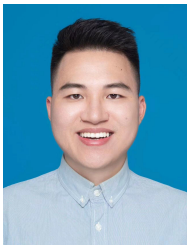
[42] Y. Ju, K.-M. Lam, W. Xie, H. Zhou, J. Dong, and B. Shi, "Deep learning methods for calibrated photometric stereo and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[43] Z. He, T. Sun, Q. Tang, K. Wang, X. Huang, and X. Qiu, "DiffusionBERT: Improving generative masked language models with diffusion models," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, 2023, pp. 4521–4534.

[44] D. Emelianenko, E. Voita, and P. Serdyukov, "Sequence modeling with unconstrained generation order," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, p. 7698–7709.
[45] D. Zhou, N. Scharli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le, and E. H. hsin Chi, "Least-to-Most prompting enables complex reasoning in large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.
[46] B. Shi, Z. Mo, Z. Wu, D. Duan, S.-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 271–284, 2019.
[47] J. Ren, F. Wang, J. Zhang, Q. Zheng, M. Ren, and B. Shi, "DiLiGenT10^2: A photometric stereo benchmark dataset with controlled shape and material variation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12 581–12 590.
[48] F. Wang, J. Ren, H. Guo, M. Ren, and B. Shi, "DiLiGenT-Pi: Photometric stereo for planar surfaces with rich details-benchmark dataset and beyond," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 9477–9487.
[49] F. Logothetis, R. Mecca, I. Budvytis, and R. Cipolla, "A cnn based approach for the point-light photometric stereo problem," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 101–120, 2023.
[50] H. Santo, M. Waechter, and Y. Matsushita, "Deep near-light photometric stereo for spatially varying reflectances," in *Proc. Int. Conf. Comput. Vis.*, 2020, pp. 137–152.
[51] D. Lichy, S. Sengupta, and D. W. Jacobs, "Fast light-weight near-field photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12 612–12 621.
[52] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Self-calibrating deep photometric stereo networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8739–8747.
[53] Z. Li, Q. Zheng, B. Shi, G. Pan, and X. Jiang, "DANI-Net: Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8381–8391.
[54] J. Li and H. Li, "Neural reflectance for shape recovery with shadow handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 221–16 230.
[55] Li, Junxuan and Li, Hongdong, "Self-calibrating photometric stereo by neural inverse rendering," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 166–183.
[56] T. Taniai and T. Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4857–4866.
[57] W. Yang, G. Chen, C. Chen, Z. Chen, and K.-Y. K. Wong, "S^3-NeRF: Neural reflectance field from shading and shadow under a single viewpoint," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1568–1582.
[58] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.
[59] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4905–4913.
[60] G. Chen, M. Waechter, B. Shi, K.-Y. K. Wong, and Y. Matsushita, "What is learned in deep uncalibrated photometric stereo?" in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 745–762.
[61] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[62] B. Li, Y. Xia, S. Xie, L. Wu, and T. Qin, "Distance-enhanced graph neural network for link prediction," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2021.
[63] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is not all you need: Pure attention loses rank doubly exponentially with depth," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2793–2803.
[64] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[65] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 22 243–22 255.
[66] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 9620–9629.
[67] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 759–769, 2003.
[68] S. Ikehata and K. Aizawa, "Photometric stereo using constrained bivariate regression for general isotropic surfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2179–2186.
[69] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep photometric stereo network," in *Proc. Int. Conf. Comput. Vis. Workshop*, 2017, pp. 501–509.
[70] D. Honzátko, E. Türetken, P. Fua, and L. A. Dunbar, "Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo," in *Proc. - Int. Conf. 3D Vis., 3DV*, 2021, pp. 394–402.
[71] C. Hardy, Y. Quéau, and D. Tschumperlé, "Uni ms-ps: A multi-scale encoder-decoder transformer for universal photometric stereo," *Comput. Vis. Image Underst.*, vol. 248, p. 104093, 2024.

Xiaoyao Wei received the B.E. degree from Zhejiang University in 2019. He is currently working toward the Ph.D. degree in mechanical manufacturing and automation with Zhejiang University, Hangzhou, China. His research interests include photometric stereo, defect detection, and deep learning.

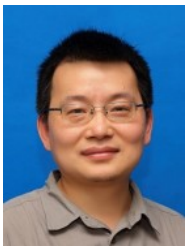
Zongrui Li received the B.E. degree from Beijing Jiaotong University, and M.S. degree from Nanyang Technological University. He is currently working toward the Ph.D. degree in Rapid-Rich Object Search (ROSE) Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include photometric stereo, NeRF, PBR, and generative model (especially for 3D generation).



Binjie Ding received the B.E. degree from Zhejiang University of Technology in 2018 and Ph.D. degree from Zhejiang University in 2023. His research interests include photometric stereo, defect detection, and deep learning.

Boxin Shi (Senior Member, IEEE) 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.

Xudong Jiang (Fellow, IEEE) received the B.E. and Ph.D. degrees in electrical engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, and the PhD degree in electrical engineering 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 CVPR/NeurIPS/ICCV/ECCV/ICLR/AAAI. 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. 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.



Gang Pan (Senior Member, IEEE) is a distinguished professor in the College of Computer Science and Technology at Zhejiang University, where he also serves as the Director of the State Key Laboratory of Brain-Machine Intelligence. He earned his B.Eng. and Ph.D. degrees from Zhejiang University in 1998 and 2004, respectively. Dr. Pan's research interests include brain-machine interfaces, brain-inspired computing, artificial intelligence, and pervasive computing. He has published more than 200 refereed publications, and has more than 60 patents granted. Dr. Pan has received numerous honors, including the NSFC Distinguished Young Scholars, the IEEE TCSC Award for Excellence (Middle Career Researcher), and the CCF-IEEE CS Young Scientist Award. Additionally, he has been recognized with the National Science and Technology Progress Award, two test-of-time paper awards, and multiple best paper awards. He serves as an associate editor for multiple prestigious journals such as IEEE Transactions on Neural Networks and Learning Systems.

Yanlong Cao received the Ph.D. degree in mechanical engineering from Zhejiang University, Hangzhou, China, in 2003. He is currently a Full-Time Professor with the State Key Laboratory of Fluid Power and Mechatronic Systems and Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China. His research interests are precision engineering, computer vision, and 3D printing.

Qian Zheng (Member, IEEE) is a tenure-track professor at the College of Computer Science and Technology, Zhejiang University. He received his B.E. (2011) and Ph.D. (2017) degrees in computer science from Zhejiang University, China. From 2018 to 2022, he served as a Research Fellow at the ROSE Lab, Nanyang Technological University. His research interests include neuromorphic computing and computer vision. He has co-authored over 60 peer-reviewed papers, with 10 papers in the IEEE journals and 30 papers in top conferences CVPR/ICCV/ECCV/NeurIPS/ICML/ICLR/AAAI. Dr. Zheng was a recipient of the ACM Rising Star Award (Hangzhou Chapter) and currently serves as an Associate Editor for journals of IEEE Transactions on Cognitive and Developmental Systems and Neurocomputing.

Supplementary Material: Revisiting Supervised Learning-Based Photometric Stereo Networks

Xiaoyao Wei, Zongrui Li, Binjie Ding, Boxin Shi, *Senior Member, IEEE,* Xudong Jiang, *Fellow, IEEE,* Gang Pan, *Senior Member, IEEE,* Yanlong Cao, and Qian Zheng, *Member, IEEE*

In this supplementary material:

1. We provide a more detailed experimental setup for validating the deep features of SL-PSNs in Section 8 (footnote 3 in the main paper); we also provide a detailed correlation analysis between the shading error of SL-PSNs [1], [2] and the normal estimation error of the L2 method [3] in Fig. 22 of this section, which further illustrates that shading constitutes the deep feature for SL-PSNs (footnote 4 in the main paper); we also provide additional qualitative analyses of deep features extracted from ID-profiles (CNN-PS [1]) and 2D-images (PS-FCN [2]) in Fig. 23 and Fig. 24 of this section, further enhancing the conclusion that shading constitutes the deep feature for SL-PSNs (footnote 5 in the main paper).
2. We present the data flow of the light-aware deep features within the proposed ESSENCE-Net in Section 9 (footnote 7 in the main paper). Changing the order of the lights does not affect the output lighting-free deep features.
3. We present the detailed performance of the proposed shading supervision strategy on 100 materials from the $SynTest^{MERL}$ [4] in Section 10 (footnote 8 in the main paper).
   - Fig. 26 (SPHERE);
   - Fig. 27 (BUNNY);
   - Fig. 28 (ARMADILLO);
   - Fig. 29 (DRAGON).
4. We present the detailed performance of the *easy-first-encoding* strategy and the effectiveness of normalization on *DiLiGenT10*^{2} dataset [5] across various materials and shapes in Section 11 (footnotes 9 and 10 in the main paper).
5. We provide a detailed discussion on the performance of models trained with different numbers of input images in both sparse and dense scenarios in Section 12 (footnote 11 in the main paper).
6. We present a qualitative comparison between the proposed ESSENCE-Net and previous state-of-the-art methods on 10 objects from the *DiLiGenT* dataset [6] with dense inputs (96 images) in Section 13 (footnote 12 in the main paper).
   - Fig. 33 (BALL, CAT, POT1);
   - Fig. 34 (BEAR, POT2, BUDDHA);
   - Fig. 35 (GOBLET, READING, COW, HARVEST).
7. We present a qualitative comparison between the proposed ESSENCE-Net and previous state-of-the-art methods on 10 objects from the *DiLiGenT* dataset [6] with sparse inputs (10 images) in Section 14 (footnote 13 in the main paper).
   - Fig. 36 (BALL, CAT, POT1);
   - Fig. 37 (BEAR, POT2, BUDDHA);
   - Fig. 38 (GOBLET, READING, COW, HARVEST).
8. We present a detailed illustration on the loss of details on the *DiLiGenT-Π* dataset in Section 15 (footnote 14 in the main paper).

- *Xiaoyao Wei, Binjie Ding, and Yanlong Cao are with the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, Hangzhou, Zhejiang 310058, China, and also with Zhejiang Key Laboratory of Advanced Equipment Manufacturing and Measurement Technology, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: {w_xy, dingbinj, sdcaoyl}@zju.edu.cn.*
- *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, 100871, China. E-mail: shiboxin@pku.edu.cn.*
- *Zongrui Li and Xudong Jiang are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798. E-mail: {zongrui001, exdjiang}@ntu.edu.sg.*
- *Qian Zheng and Gang Pan are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China, and also with the State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, Zhejiang 310027, China. E-mail: {qianzheng, gpan} @zju.edu.cn.*

Corresponding authors: Zongrui Li, Yanlong Cao and Qian Zheng

8 Detailed Experimental Setup for Validating the Deep Features of SL-PSNs

Experimental setup for per-pixel methods. For per-pixel methods, which extract deep features from 1D-profiles, we analyze the typical network, CNN-PS [1], which has become the foundation of numerous works [7], [8], [9], [10]. Since the spatial positions in the observation map correspond to the light directions, and the normal is given, we can calculate a dense ground truth shading map. We analyze the features after each ReLU (rectified linear unit) layer in *light-aware deep feature decoder* by taking the channel average. The mean absolute error between the deep features extracted by the network and ground truth shading is calculated and averaged over all points.

Experimental setup for all-pixel methods. For all-pixel methods, which extract features from 2D-images, we analyze the typical network, PS-FCN [2], whose concept has been adopted by many studies [4], [11], [12]. We take the max value of the features by channel dimension as PS-FCN [2] uses features after max-pooling layer for normal estimation. We then calculate the error between the feature map and ground truth shading in *light-aware deep feature decoder*. Note that the deconvolution operation will introduce a checkerboard pattern [13] and seriously affect the analysis. Thus, we only analyze the features before the deconvolution operation.

Correlation analysis between shading error of SL-PSNs and normal estimation error of the L2 method. Fig. 22 shows a strong correlation between shading error of the deepest layer (Feature 5) extracted by representative SL-PSNs and normal error estimated by L2 method [3]. Since the L2 method directly recovers normals from shading, the normal estimation error inversely reflects the deviation from shading. Therefore, such a strong correlation manifests that the shading error we compute is highly correlated with the actual deviation from the ground truth shading, which enhances the validity of our quantitative verification (shading constitutes the deep feature for SL-PSNs).

Qualitative analyses of deep features extracted by CNN-PS [1] from ID-profiles and PS-FCN [2] from 2D-images. For CNN-PS [1], which extracts deep features from 1D-profiles, we randomly sampled an observation point on the object from the *DiLiGenT* dataset [6] and visualized the extracted deep features and the ground truth shading in Fig.23. As the network goes deeper, the extracted deep features become increasingly continuous and closer to the ground truth shading. For PS-FCN [2], which extracts deep features from 2D-images, we randomly sampled an observed image under one lighting direction on the object from the *DiLiGenT* dataset [6] and displayed the extracted deep features and the ground truth shading in Fig. 24. As the feature extraction process goes deeper, the extracted deep features become increasingly similar to the ground true shading. These experiments further enhance our conclusion that shading constitutes the light-aware deep feature extracted by SL-PSNs.

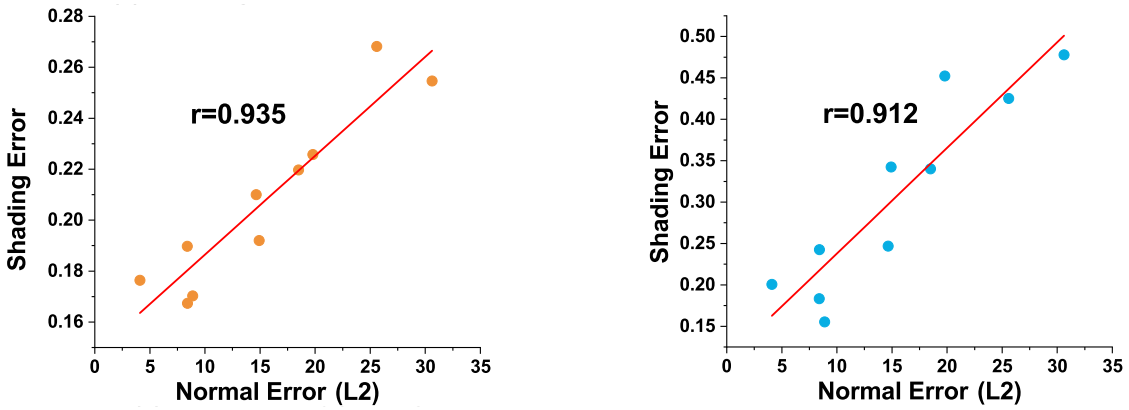


Fig. 22. Correlation analysis between shading error (the deepest layer) of SL-PSNs and normal estimation error of the L2 method [3]. 'r' represents the correlation coefficient. (a) Correlation between shading error of per-pixel method CNN-PS [1] and normal estimation error of the L2 method [3]. (b) Correlation between shading error of all-pixel method PS-FCN [2] and normal estimation error of the L2 method [3].

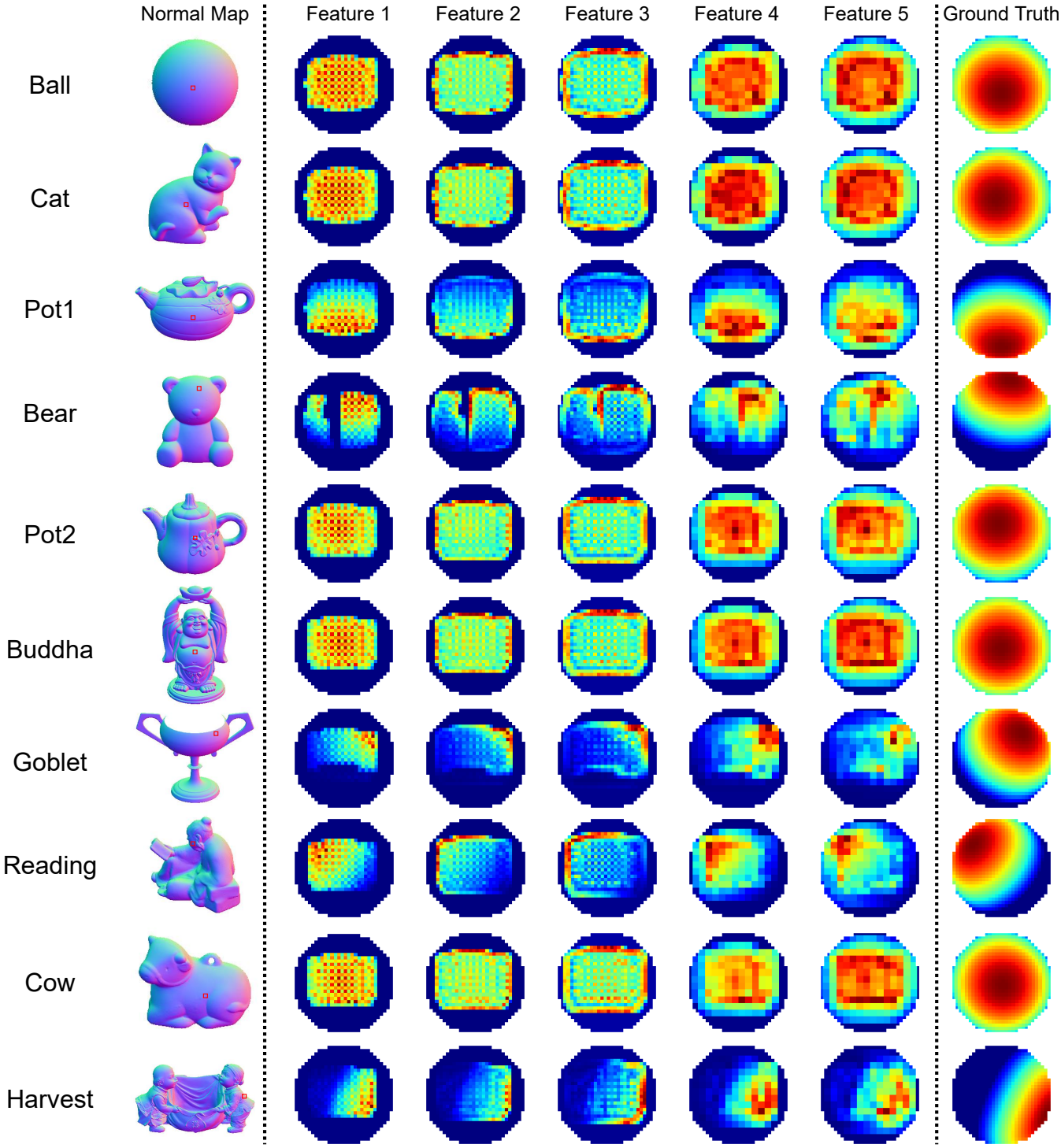


Fig. 23. Qualitative analysis of the relationship between the deep features extracted from 1D-profiles by CNN-PS [1] and the ground truth shading. The leftmost column shows the normal map. The middle five columns display the deep features extracted by the first five layers in *light-aware deep feature encoder*, corresponding to the point marked by the red box in the leftmost column. The last column shows the ground truth shading.

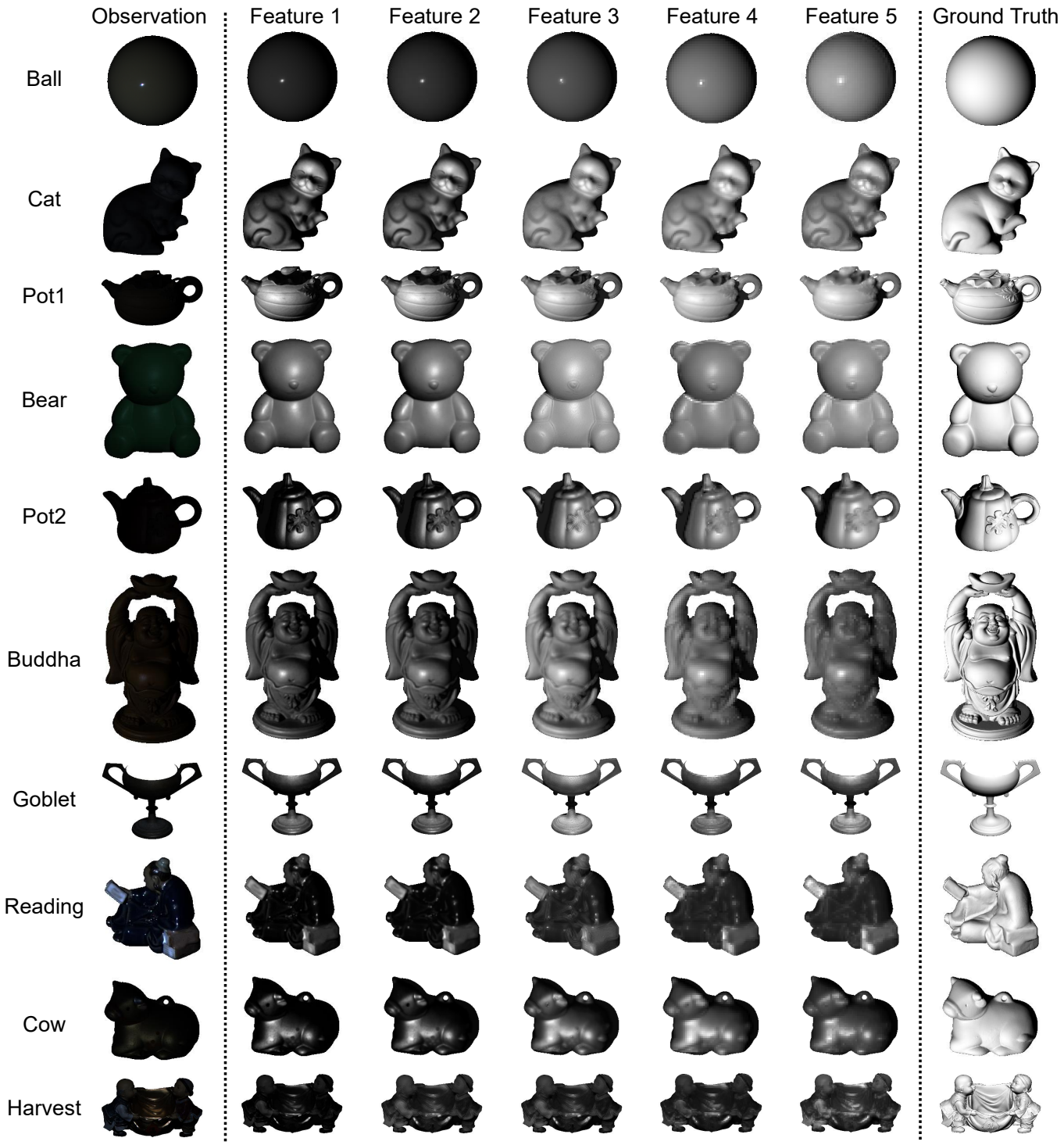


Fig. 24. Qualitative analysis of the relationship between the deep features extracted from 2D-images by PS-FCN [2] and the ground truth shading. The leftmost column shows the observed image under one random light. The middle five columns display the deep features extracted by the first five layers in *light-aware deep feature encoder*, corresponding to the observed image in the leftmost column. The last column shows the ground truth shading.

9 The Data Flow of Light-Aware Deep Features

In Fig. 25, we show the data flow of light-aware deep features within ESSENCE-Net. When the order of the input image and light sequence changes, the output light-free deep features remain identical. We further illustrate this in terms of specific components in ESSENCE-Net:

- **Convolutions for dimensionality increase and spatial downsampling (conv1 and conv2):** The operations are performed on 2D-images and do not involve the 1D-profiles (light dimension), where the convolutional weights are shared across different light directions.
- **Transformer for feature encoding from 1D-profiles (1DP-FE):** The self-attention mechanisms across the light dimension ensure invariance to the order of lights [14].
- **CNN for feature encoding from 2D-images (2DI-FE):** The convolution operation is performed on 2D-images for feature extraction without involving the light dimension, where the convolutional weights are shared across different light directions.
- **Max-pooling for obtaining light-free deep features:**The maximum values from the 1D-profile dimension are preserved, which are inherently unaffected by the order of lights [2], [4].
- **Normal decoding:** Since light information has been removed, the decoding of normals is inherently independent of the light order.

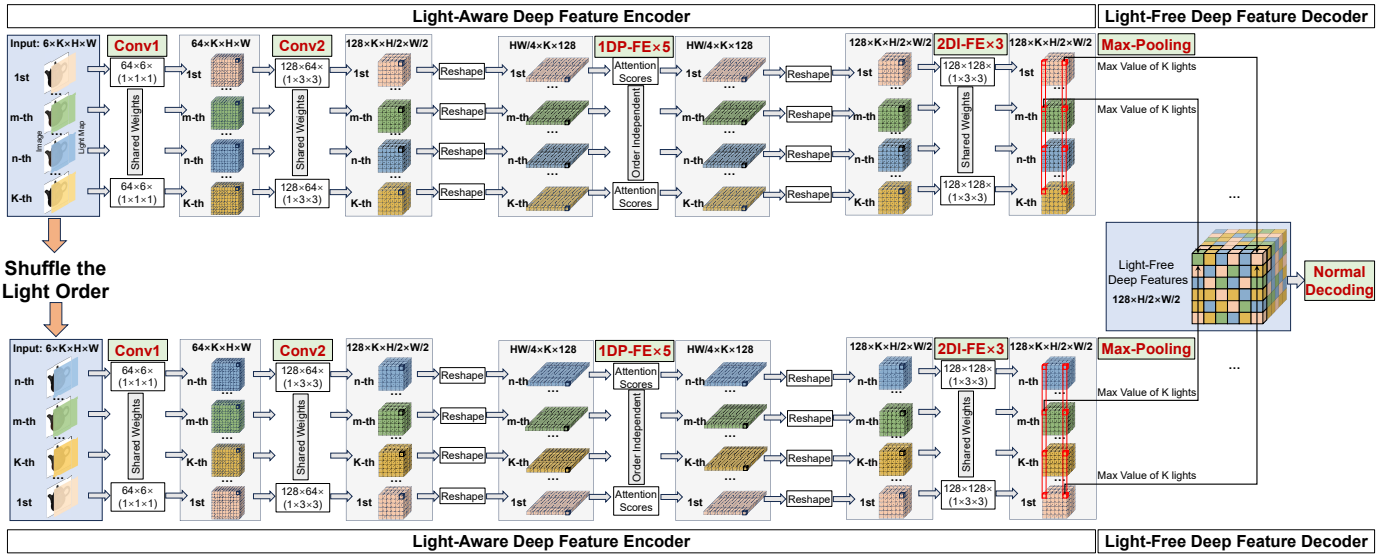


Fig. 25. Data flow of light-aware deep features within ESSENCE-Net. Changes in the order of lights do not affect the extracted light-free deep features.

10 Effectiveness of Shading Supervision Strategy on Data with Different Materials

From Fig. 26 to Fig. 29, we show the performance of the proposed method *w/* and *w/o* shading supervision on $SynTest^{MERL}$ dataset [4], which includes SPHERE, BUNNY, ARMADILLO, and DRAGON objects, each with 100 materials from the *MERL* dataset [15]. The figures illustrate that the proposed method outperforms the method *w/o* shading supervision on objects of almost all materials. This demonstrates the effectiveness of the shading supervision strategy across various materials, whether close to diffuse materials or non-diffuse materials.

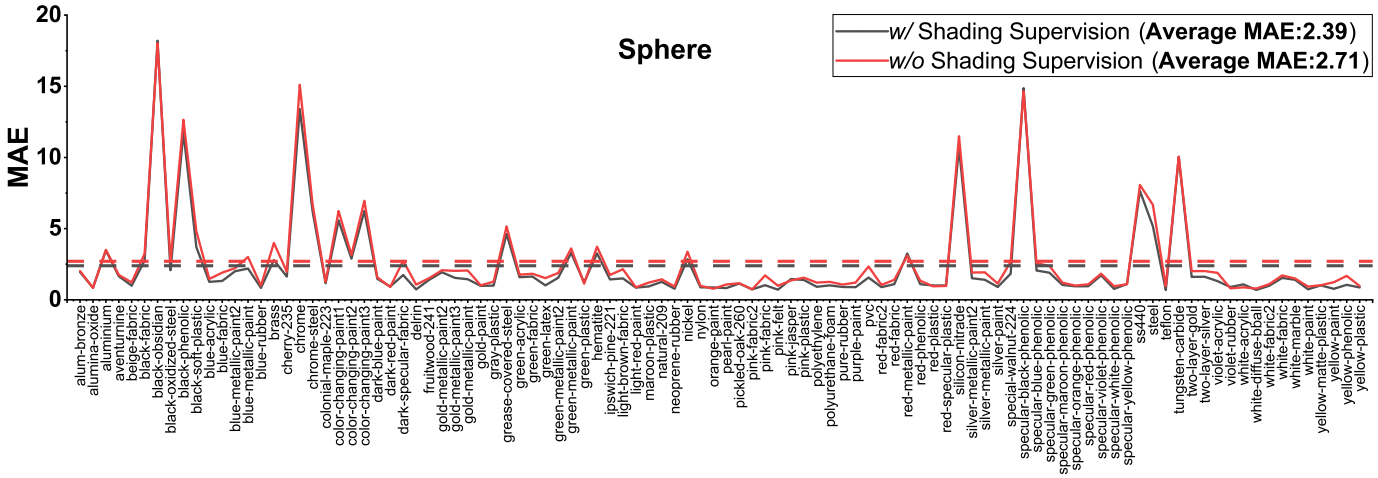


Fig. 26. The performance of the proposed method *w/* and *w/o* shading supervision on SPHERE from $SynTest^{MERL}$ dataset [4]. The red and gray dashed lines represent the average MAE of methods *w/* and *w/o* shading supervision across the 100 materials.

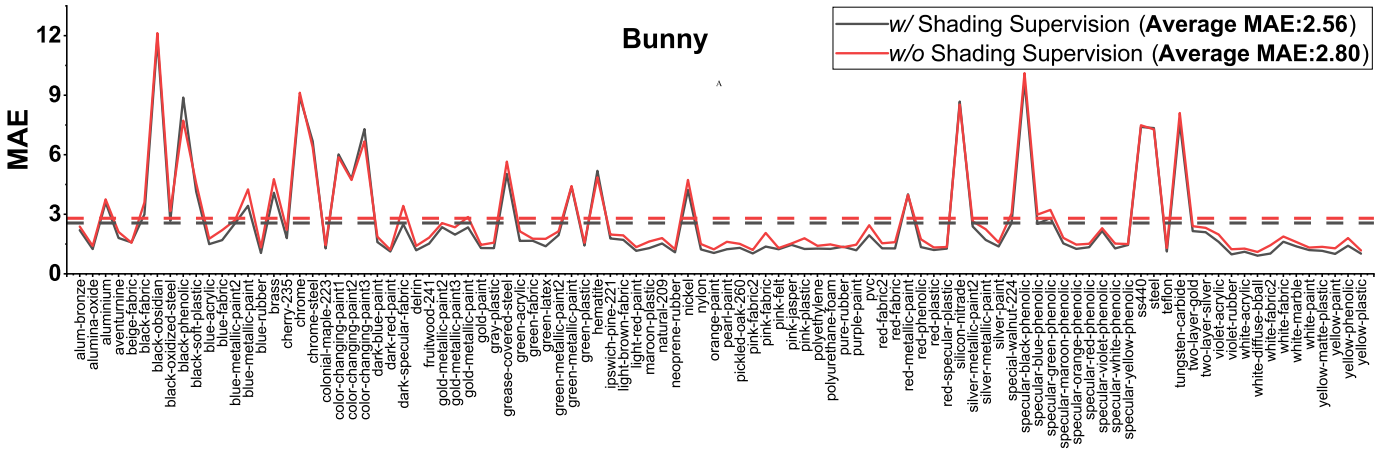


Fig. 27. The performance of the proposed method *w/* and *w/o* shading supervision on BUNNY from $SynTest^{MERL}$ dataset. The red and gray dashed lines represent the average MAE of methods *w/* and *w/o* shading supervision across the 100 materials.

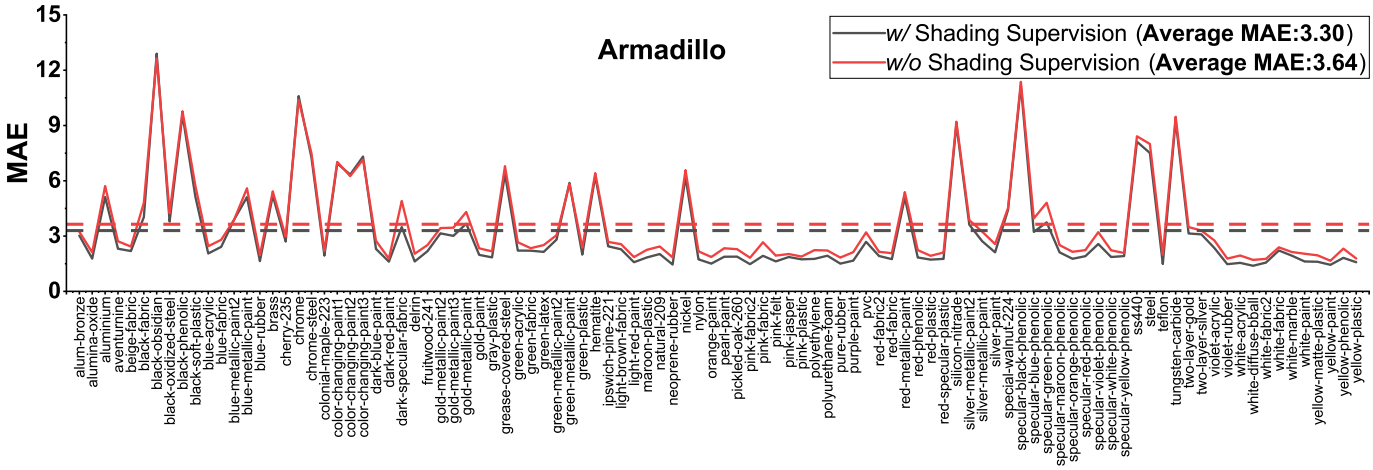


Fig. 28. The performance of the proposed method *w/* and *w/o* shading supervision on ARMADILLO from $SynTest^{MERL}$ dataset [4]. The red and gray dashed lines represent the average MAE of methods *w/* and *w/o* shading supervision across the 100 materials.

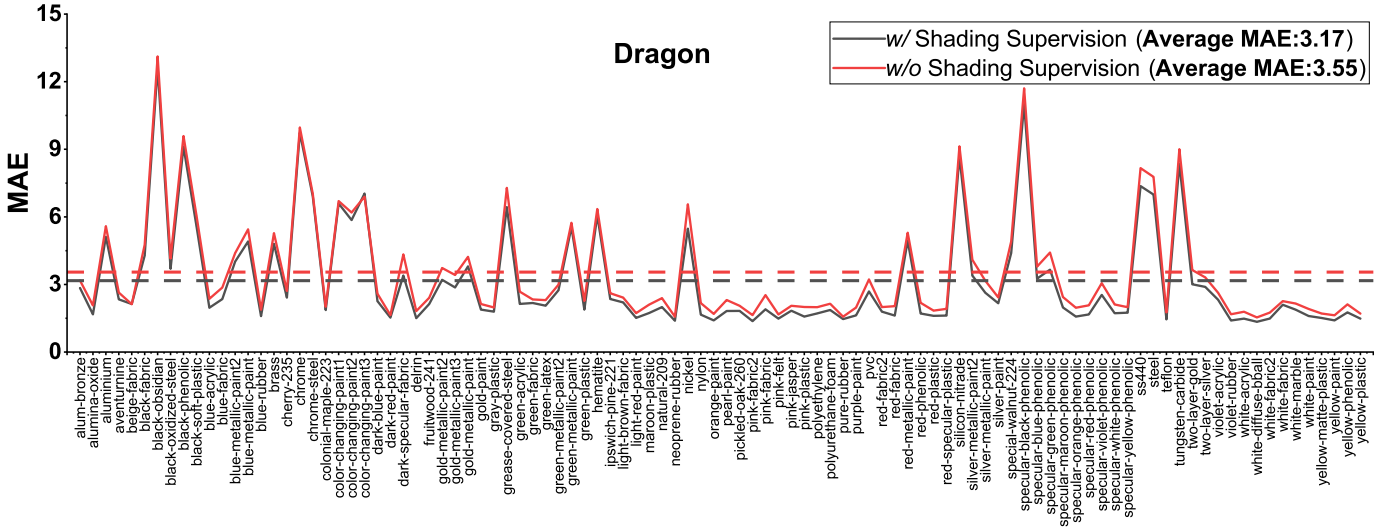
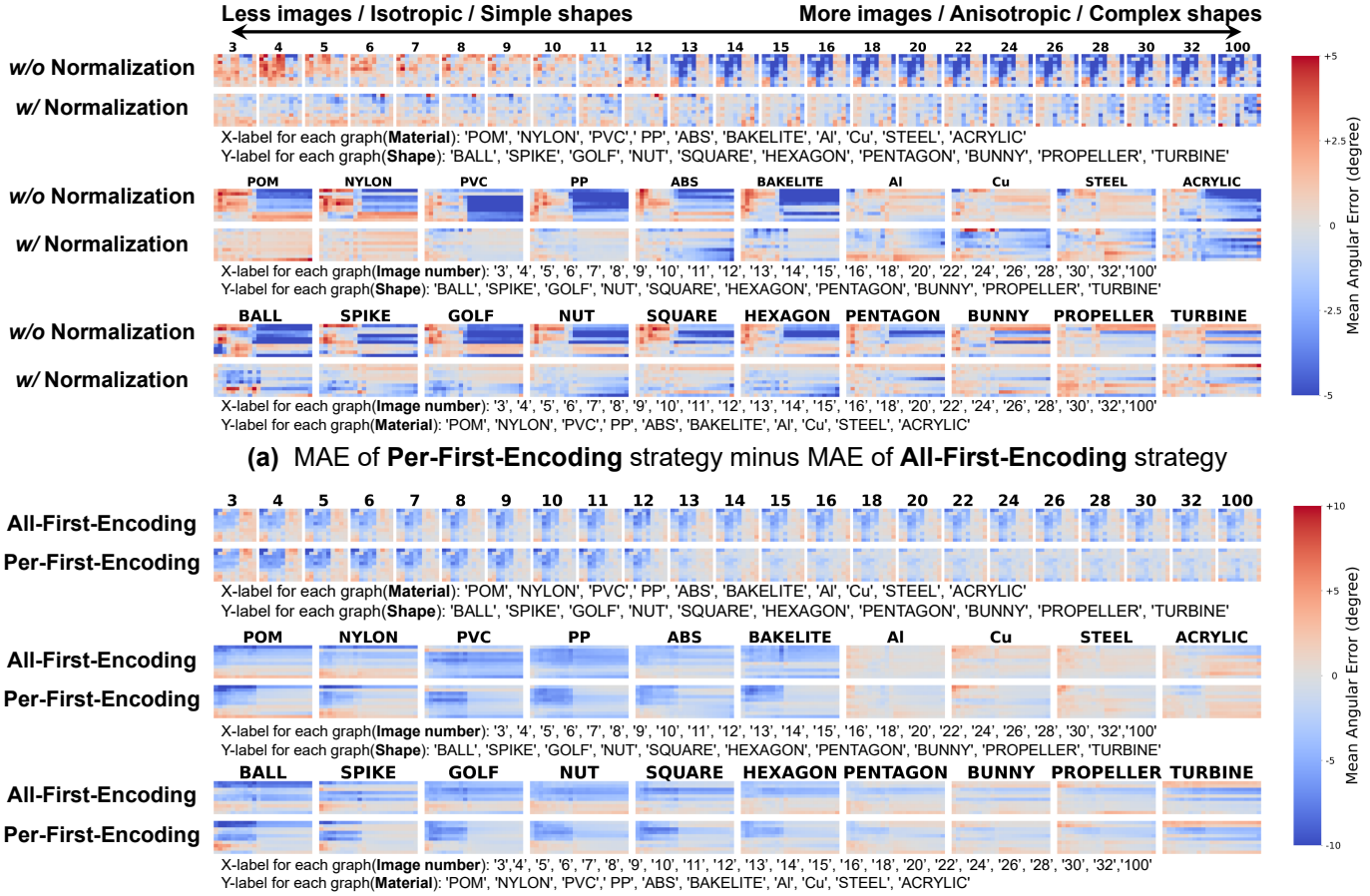


Fig. 29. The performance of the proposed method *w/* and *w/o* shading supervision on DRAGON from $SynTest^{MERL}$ dataset [4]. The red and gray dashed lines represent the average MAE of methods *w/* and *w/o* shading supervision across the 100 materials.

11 Effectiveness of Easy-First-Encoding and Normalization Strategy on DiLiGenT10^2 Dataset

Fig. 30a illustrates the performance of the proposed *easy-first-encoding* strategy on *DiLiGenT10*^{2} dataset [5], across various materials and shapes. The per-first-encoding strategy gradually surpasses the all-first-encoding strategy, demonstrating the effectiveness of the proposed *easy-first-encoding* strategy, as encoding deep features from 1D-profiles becomes easier with an increasing number of images. And the use of normalization enhances this advantage even further. Fig. 30b illustrates the performance of the normalization strategy on *DiLiGenT10*^{2} dataset [5] across various materials and shapes, showing significant effects on isotropic and simple-shaped surfaces while still beneficial but less pronounced for anisotropic and complex-shaped surfaces.



(a) MAE of **Per-First-Encoding** strategy minus MAE of **All-First-Encoding** strategy

(b) MAE of ***w/* Normalization** minus MAE of ***w/o* Normalization**

Fig. 30. The detailed performance of the *easy-first-encoding* strategy and the effectiveness of normalization on *DiLiGenT10*^{2} dataset [5] across various materials and shapes. Each subplot in the figure represents performance on a specific image number, shape, or material, with the x-axis as the horizontal axis and the y-axis as the vertical axis.

12 Performance of Models Trained with Sparse and Dense Inputs in Spare Test Scenarios

In Fig. 31, we present the average MAE (AMAE) of normal maps estimated by models trained with sparse (10) and dense (32) inputs, given sparse (10) inputs during testing. All results are reported as the average of 100 random experiments. Specifically, slight performance degradation in AMAE is observed when comparing models trained with dense input ($AMAE^{32}$) to sparse input ($AMAE^{10}$) (*e.g.*, $1.15°$ on the *DiLiGenT* dataset [6], $0.14°$ on the *DiLiGenT10²* dataset [5], and $2.05°$ on the *DiLiGenT-Π* dataset [16], $AMAE^{32}$ – $AMAE^{10}$), which is inevitable due to the gap in shading features extracted by the deep network during the testing and training stages. This degradation is particularly evident on anisotropic objects, where the extracted pattern may be more sensitive to the number of input images due to the complexity of its reflectance model [17], [18].

We further compare our method with PS-FCN^{+N} [4][15], as shown in Fig. 32. The comparison reveals improved stability in our method while maintaining an advantage in normal estimation accuracy. This improvement is particularly evident for anisotropic material, which highlights the effectiveness of the *easy-first-encoding* strategy in addressing PS challenges in complex scenarios. It also underscores the strength of the network design, particularly in using optimal network structure to extract limited but useful information in 1D-profiles and 2D-images.

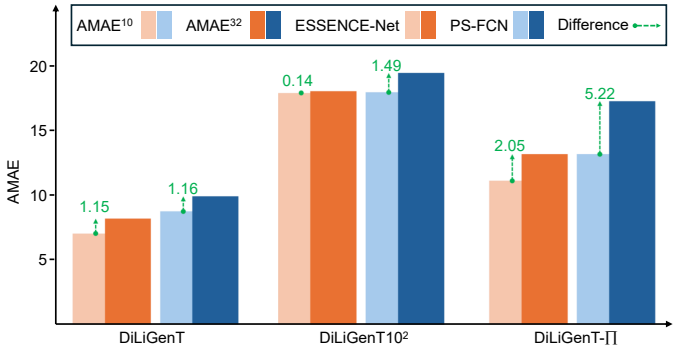


Fig. 31. Quantitative analysis on ESSENCE-Net and PS-FCN^{+N} [4] trained with 10 images and 32 images, tested with 10 images. The average MAE (AMAE) over 100 random experiments are calculated for models trained with 10 images ($AMAE^{10}$) and 32 images ($AMAE^{32}$) across three datasets: *DiLiGenT* [6], *DiLiGenT10²* [5], and *DiLiGenT-Π* [16]. The green dotted line and values represent the difference, $AMAE^{32}$ - $AMAE^{10}$.

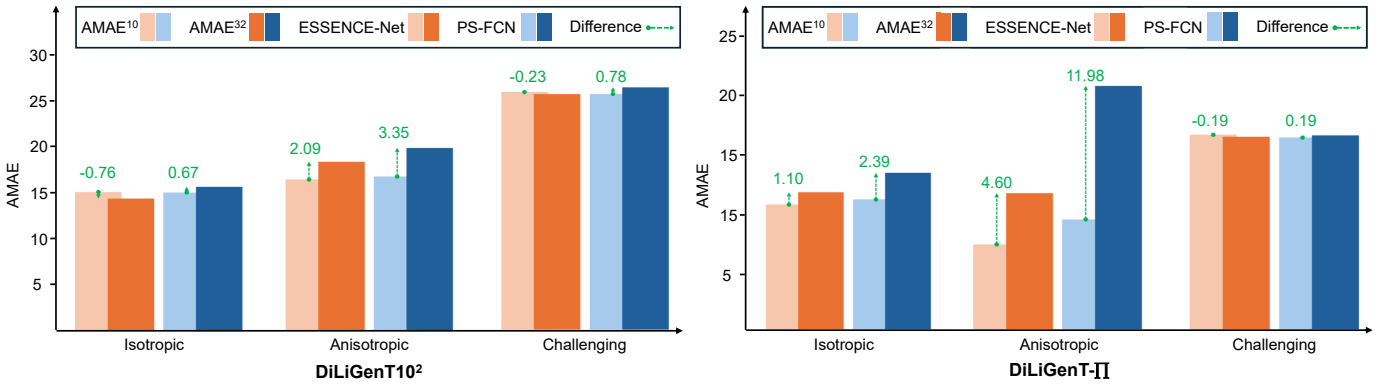


Fig. 32. Quantitative analysis on ESSENCE-Net and PS-FCN^{+N} [4] trained with 10 images and 32 images, tested with 10 images. The average average MAE (AMAE) over 100 random experiments are calculated for models trained with 10 images ($AMAE^{10}$) and 32 images ($AMAE^{32}$) across three groups of materials, *i.e.*, isotropic, anisotropic, and challenging (translucent) on *DiLiGenT10²* [5] and *DiLiGenT-Π* [16] datasets. The green dotted line and values represent the difference, $AMAE^{32}$ - $AMAE^{10}$.

15. We choose PS-FCN^{+N} [4] as the baseline method since it is one of the most representative, open-source SL-PSN that shares the same training setting as ESSENSE-Net and achieves relatively low normal estimation errors. Methods with hybrid encoder structures are not included in the comparison as they do not release their training code.

13 Qualitative Analysis on *DiLiGenT* Dataset with Dense Inputs

From Fig. 33 to Fig. 35, we show a comparison between ESSENCE-Net and previous state-of-the-art methods on 10 objects from *DiLiGenT* dataset [6], in terms of the estimated normal maps and error maps. The comparison includes conventional method (ST14 [17]), per-pixel method (CNN-PS [1]), all-pixel method (PS-FCN^{+N} [4]), and hybrid method (GPS-Net [19]). ESSENCE-Net outperforms in general reflective surfaces (*e.g.*, CAT and COW) and complex surfaces heavily influenced by global illumination effects (*e.g.*, READING and HARVEST), indicating the effectiveness of the proposed ESSENCE-Net in handling the challenges posed by unknown reflectance and global illumination effects.

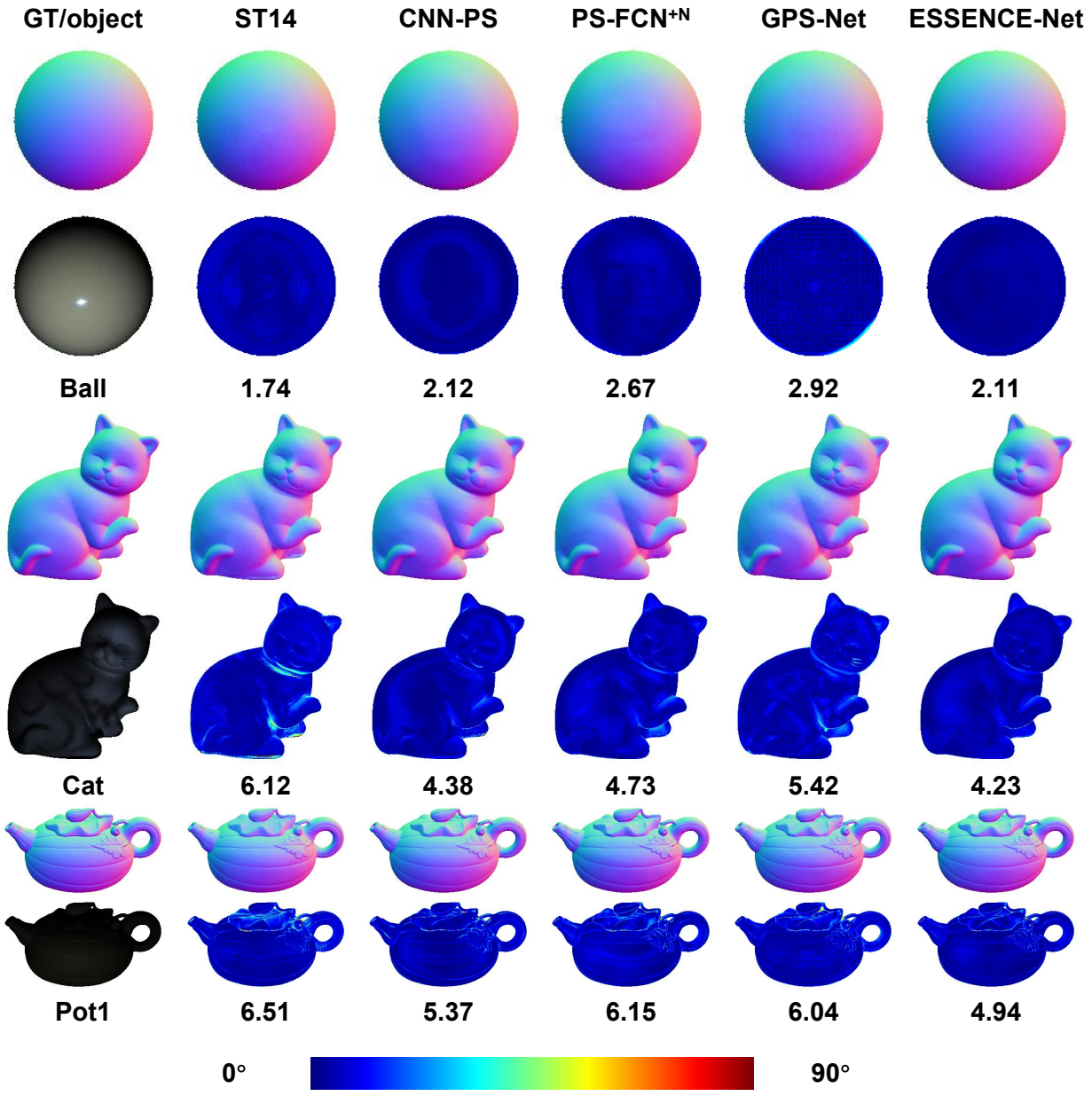


Fig. 33. Comparison of the proposed ESSENCE-Net with existing state-of-the-art methods (ST14 [17], CNN-PS [1], PS-FCN^{+N} [4], GPS-Net [19]) on estimated normal and error maps for the BALL, CAT, and POT1 objects with 96 input images. The number indicate the Mean Angular Error (MAE) for estimated normal maps.

Fig. 34. Comparison of the proposed ESSENCE-Net with existing state-of-the-art methods (ST14 [17], CNN-PS [1], PS-FCN^{+N} [4], GPS-Net [19]) on estimated normal and error maps for the BEAR, POT2, and BUDDHA objects with 96 input images. The number indicate the Mean Angular Error (MAE) for estimated normal maps.

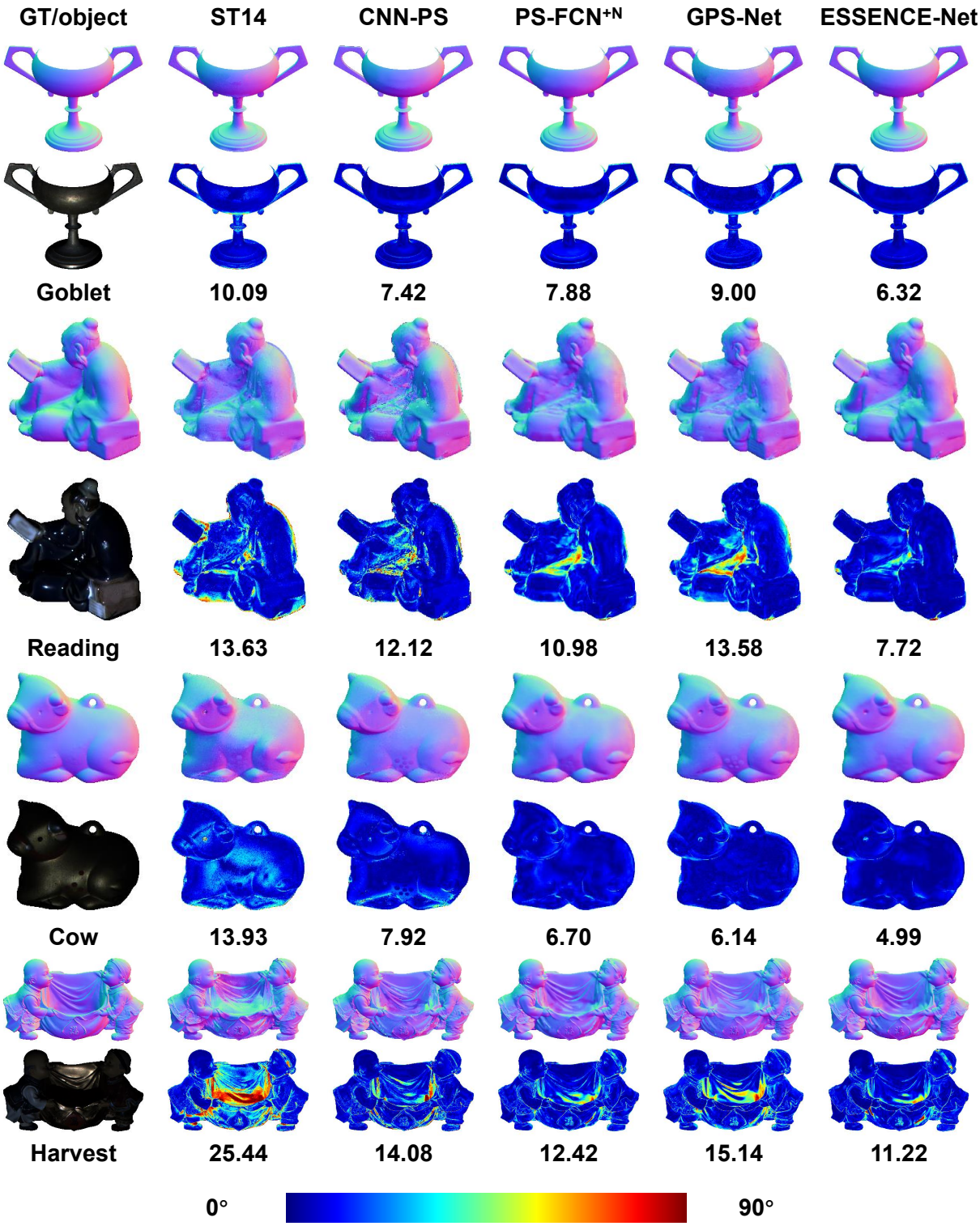


Fig. 35. Comparison of the proposed ESSENCE-Net with existing state-of-the-art methods (ST14 [17], CNN-PS [1], PS-FCN^{+N} [4], GPS-Net [19]) on estimated normal and error maps for the GOBLET, READING, COW, and HARVEST objects with 96 input images. The number indicate the Mean Angular Error (MAE) for estimated normal maps.

14 Qualitative Analysis on *DiLiGenT* Dataset with Sparse Inputs

Fig. 36 to Fig. 38 compares the performance of the proposed ESSENCE-Net with state-of-the-art sparse photometric stereo methods on BALL, CAT, and READING in terms of normal and error maps with 10 images. The comparative methods include LMPS [7], PS-Transformer [20], and SPLINE-Net [10], all designed specifically for sparse settings. For SPLINE-Net [10], PS-Transformer [20] and the proposed ESSENCE-Net, We conducted 100 random experiments and selected the normal maps and error maps closest to the average MAE for qualitative analysis. For LMPS [7], which applies a connection table to select helpful light for normal recovery, we obtained the normal maps and error maps using the released code. It is evident that our method performs exceptionally well under sparse inputs, especially in shadow and highlight areas, showing significant superiority over existing methods. The experimental results indicate that even under sparse setups, the proposed ESSENCE-Net effectively handles unknown reflectance and global illumination effects.

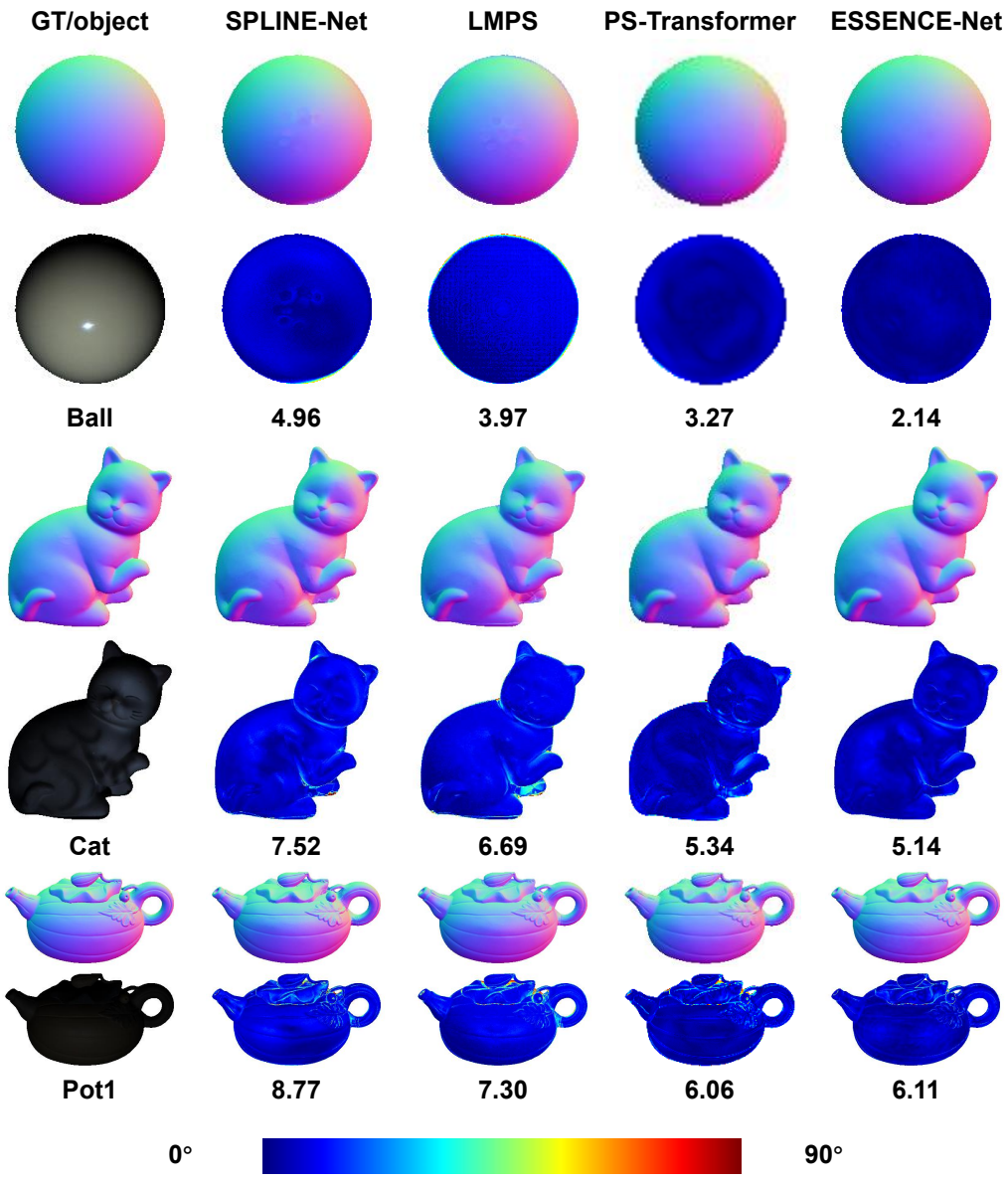


Fig. 36. Comparison of the proposed ESSENCE-Net with existing state-of-the-art methods (SPLINE-Net [10], LMPS [7], and PS-Transformer [20]) on estimated normal and error maps for the BALL, CAT, and POT1 objects with 10 input images. The comparison methods are all specifically designed for sparse inputs. For SPLINE-Net [10], PS-Transformer [20] and the proposed ESSENCE-Net, We conducted 100 random experiments and selected the normal maps and error maps closest to the average MAE for qualitative analysis, the numbers represent the average MAE of 100 random experiments. For LMPS [7], which applies a connection table to select helpful light for normal recovery, we obtained the normal maps and error maps using the released code.

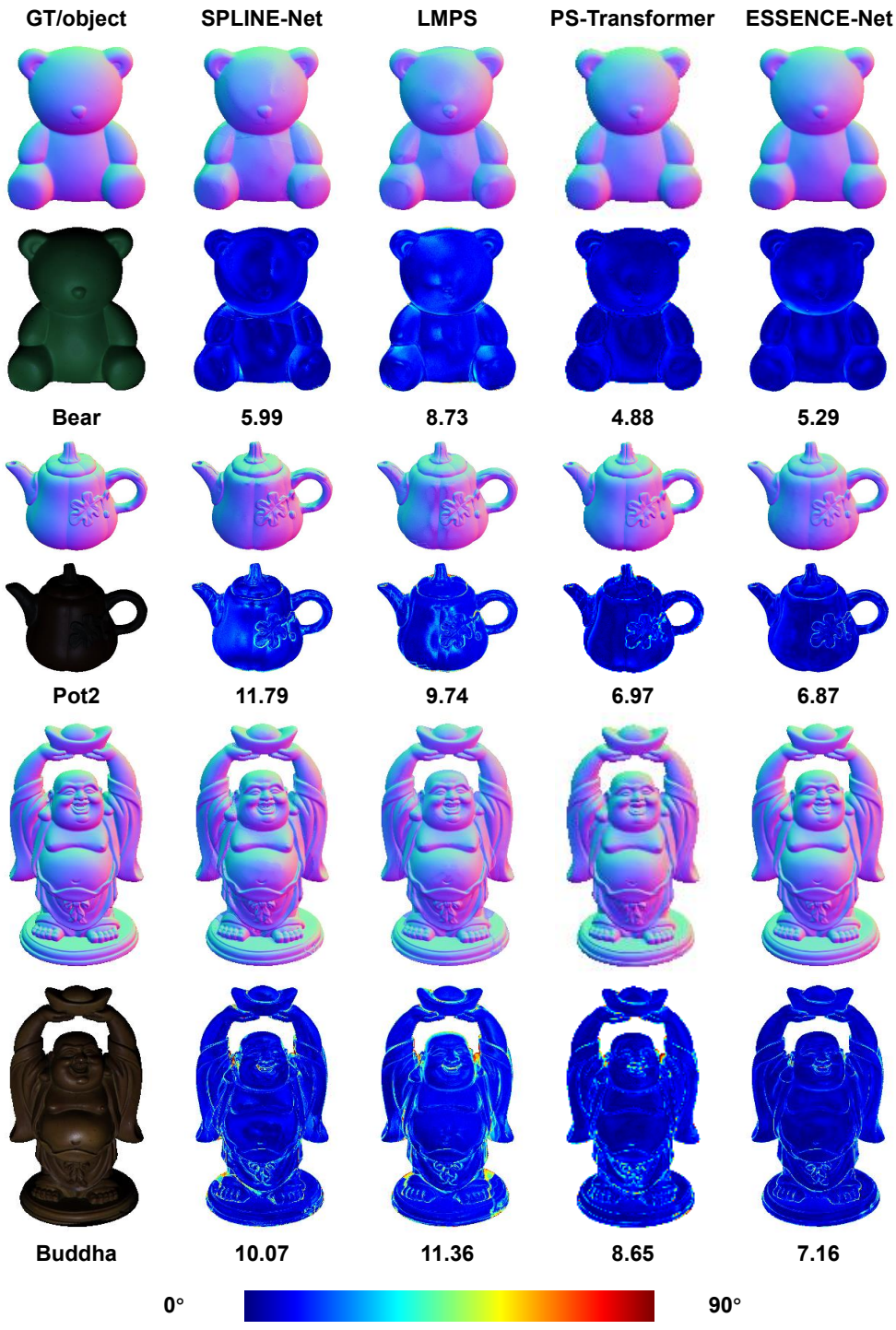


Fig. 37. Comparison of the proposed ESSENCE-Net with existing state-of-the-art methods (SPLINE-Net [10], LMPS [7], and PS-Transformer [20]) on estimated normal and error maps for the BEAR, POT2, and BUDDHA objects with 10 input images. The comparison methods are all specifically designed for sparse inputs. For SPLINE-Net [10], PS-Transformer [20] and the proposed ESSENCE-Net, We conducted 100 random experiments and selected the normal maps and error maps closest to the average MAE for qualitative analysis, the numbers represent the average MAE of 100 random experiments. For LMPS [7], which applies a connection table to select helpful light for normal recovery, we obtained the normal maps and error maps using the released code.

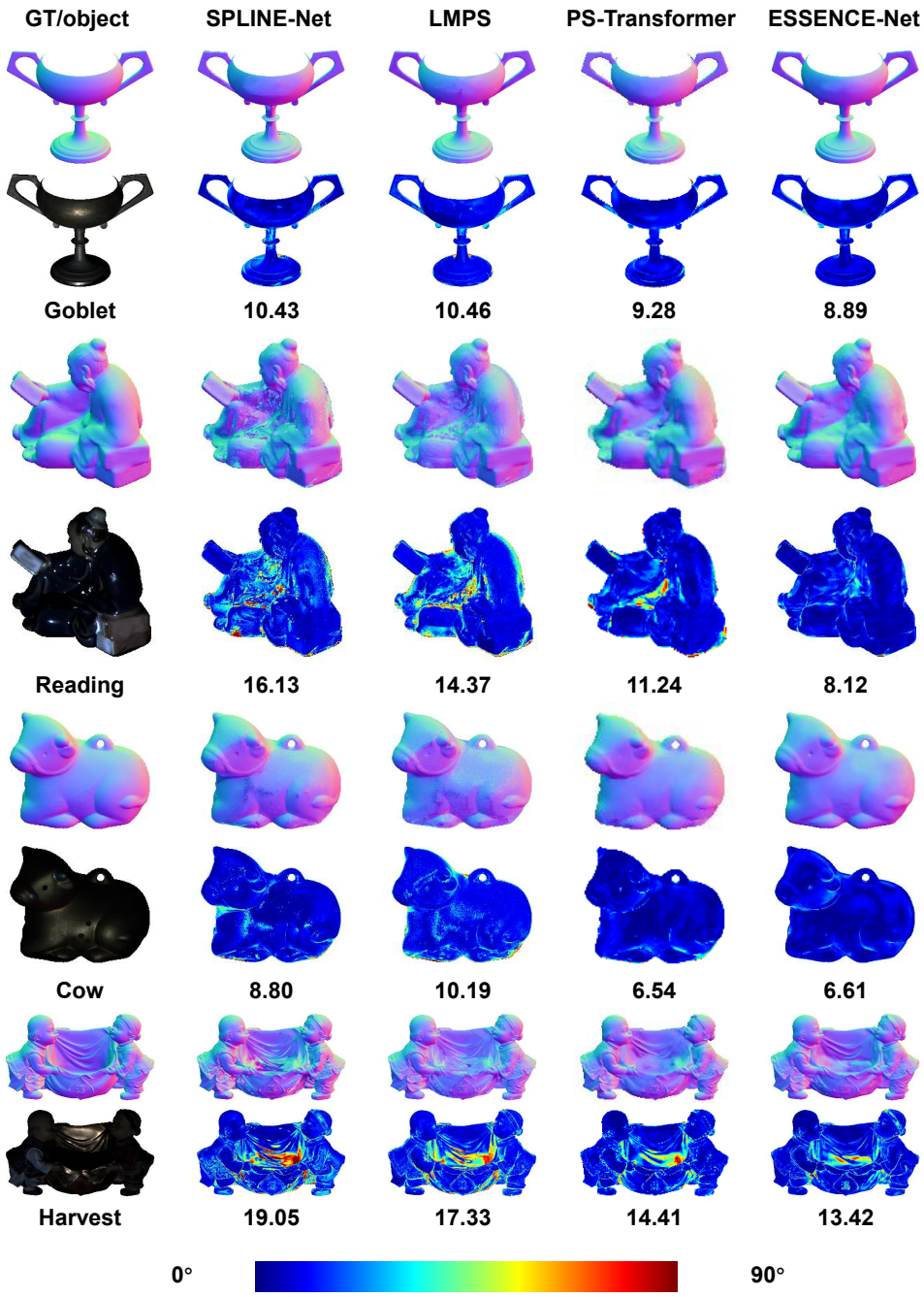


Fig. 38. Comparison of the proposed ESSENCE-Net with existing state-of-the-art methods (SPLINE-Net [10], LMPS [7], and PS-Transformer [20]) on estimated normal and error maps for the GOBLET, READING, COW, and HARVEST with 10 input images. The comparison methods are all specifically designed for sparse inputs. For SPLINE-Net [10], PS-Transformer [20] and the proposed ESSENCE-Net, We conducted 100 random experiments and selected the normal maps and error maps closest to the average MAE for qualitative analysis, the numbers represent the average MAE of 100 random experiments. For LMPS [7], which applies a connection table to select helpful light for normal recovery, we obtained the normal maps and error maps using the released code.

15 Illustration on the Loss of Details in *DiLiGenT-Π* Dataset

Compared to *DiLiGenT* [6] and *DiLiGenT10²* [5] datasets, the *DiLiGenT-Π* [16] dataset contains richer details and higher resolution, with neighboring pixels often exhibiting distinct features, as shown in Fig. 39. For per-pixel methods such as L2 [3] and CNN-PS [1], no spatial downsampling or upsampling is performed, preserving the detailed information. In contrast, all-pixel methods like PS-FCN [2] employ two downsampling and upsampling operations, leading to a loss of details that can compromise the accuracy of normal estimation. To broaden the receptive field and achieve a lightweight model, ESSENCE-Net adopts a single downsampling operation and restores the size through upsampling. As shown in Fig. 39, the details of the CLOUD-T object within the red box are magnified for comparison. The normal maps estimated by L2 [3] and CNN-PS [1] retain rich details, whereas the normal map estimated by PS-FCN [2] exhibits a noticeable loss of details. While the detailed representation in ESSENCE-Net's estimated normal map does not perfectly align with the ground truth, it still demonstrates performance comparable to CNN-PS [1]. We believe that in the future, removing the downsampling and upsampling operations on high-performance devices or integrating super-resolution techniques [21], [22], [23] could further enhance its performance.

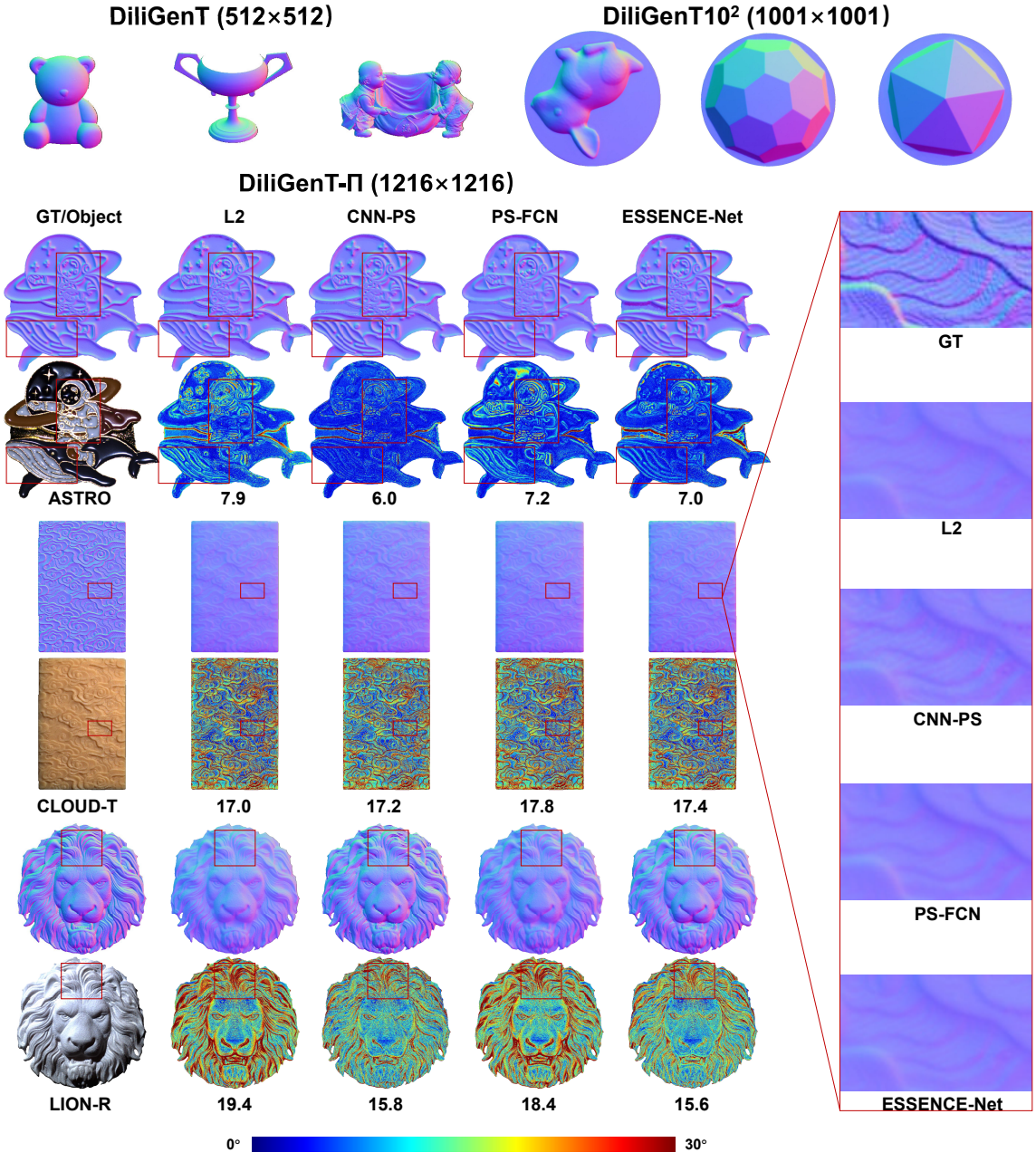


Fig. 39. A visual comparison between methods involving spatial downsampling and upsampling (PS-FCN [2], the proposed ESSENCE-Net) and methods that do not involve spatial downsampling and upsampling (L2 [3], CNN-PS [1]) on the *DiLiGenT-Π* dataset [16]. Methods of L2 and CNN-PS demonstrate better detail preservation. In contrast, PS-FCN smooths some of the details. Despite undergoing downsampling and upsampling, the proposed ESSENCE-Net still achieves excellent detail representation.

REFERENCES

[1] S. Ikehata, "CNN-PS: CNN-based photometric stereo for general non-convex surfaces," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.
[2] G. Chen, K. Han, and K.-Y. K. Wong, "PS-FCN: A flexible learning framework for photometric stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.
[3] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Opt. Eng.*, vol. 19, no. 1, pp. 139–144, 1980.
[4] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Deep photometric stereo for non-lambertian surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 129–142, 2020.
[5] J. Ren, F. Wang, J. Zhang, Q. Zheng, M. Ren, and B. Shi, "DiLiGenT10^2: A photometric stereo benchmark dataset with controlled shape and material variation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12 581–12 590.
[6] B. Shi, Z. Mo, Z. Wu, D. Duan, S.-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 271–284, 2019.
[7] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita, "Learning to minify photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7568–7576.
[8] Ikehata, Satoshi, "Does physical interpretability of observation map improve photometric stereo networks?" in *ICIP*, 2022, pp. 291–295.
[9] F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla, "PX-NET: Simple and efficient pixel-wise training of photometric stereo networks," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 12 757–12 766.
[10] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L.-Y. Duan, and A. C. Kot, "SPLINE-Net: Sparse photometric stereo through lighting interpolation and normal estimation networks," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8549–8558.
[11] G. Chen, M. Waechter, B. Shi, K.-Y. K. Wong, and Y. Matsushita, "What is learned in deep uncalibrated photometric stereo?" in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 745–762.
[12] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Self-calibrating deep photometric stereo networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8739–8747.
[13] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.
[14] H. Liu, Y. Yan, K. Song, and H. Yu, "SPS-Net: Self-attention photometric stereo network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2020.
[15] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 759–769, 2003.
[16] F. Wang, J. Ren, H. Guo, M. Ren, and B. Shi, "DiLiGenT-Pi: Photometric stereo for planar surfaces with rich details-benchmark dataset and beyond," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 9477–9487.
[17] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1078–1091, 2013.
[18] M. Holroyd, J. Lawrence, G. Humphreys, and T. Zickler, "A photometric approach for estimating normals and tangents," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 1–9, 2008.
[19] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi, "GPS-Net: Graph-based photometric stereo network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 10 306–10 316.
[20] Ikehata, Satoshi, "PS-Transformer: Learning sparse photometric stereo network using self-attention mechanism," in *Proc. Brit. Mach. Vis. Conf.*, 2021, pp. 1–12.
[21] Y. Ju, M. Jian, C. Wang, C. Zhang, J. Dong, and K.-M. Lam, "Estimating high-resolution surface normals via low-resolution photometric stereo images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2512–2524, 2024.
[22] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, 2020.
[23] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2015.