Bridge-UniPS: Bridging Calibrated Photometric Stereo toward Universal Photometric Stereo

Minzhe Xu[1], Xiaoyan Liu[2] *, Yujie Xing[1] *, and Qian Chen[1] *

[1] Huazhong University of Science and Technology, Wuhan 430074, China
[2] The Chinese University of Hong Kong, Shatin, Hong Kong, China
liuxy185@link.cuhk.edu.hk

Abstract. Universal photometric stereo (UniPS) aims to recover surface normals from images captured under unconstrained illumination, which fundamentally differs from calibrated photometric stereo (CPS), where lighting conditions are known. This gap forces most existing UniPS methods to adopt fully end-to-end learning that struggles to disentangle illumination and surface geometry due to the lack of explicit physical constraints. In this paper, we propose **Bridge-UniPS**, a novel framework that bridges CPS and UniPS through a lighting adapter trained without direct intermediate supervision. The adapter transforms images captured under unconstrained illumination into a sequence of image–illumination pairs directly compatible with CPS models, allowing CPS to act as a physics-aligned intermediate representation and thereby improving the accuracy of surface normal recovery. The lighting adapter receives no direct supervision on Bridge Images or target lighting directions, but is optimized through the final normal-estimation objective with gradients propagated through a frozen CPS network. It can still produce CPS-compatible Bridge Images while preserving illumination-geometry consistency. This behavior suggests that modern learning-based CPS models provide useful physical priors to guide the reconstruction. Extensive experiments on multiple benchmarks demonstrate that Bridge-UniPS achieves state-of-the-art performance under unconstrained illumination, exhibiting strong generalization and significantly improving fine-scale surface normal accuracy.

Keywords: Photometric Stereo, Normal Estimation, Deep Learning, Intermediate Representation

1 Introduction

Photometric stereo (PS) estimates pixel-wise surface normals from images captured under varying illumination [23,43]. Traditional Calibrated PS (CPS) [8,20, 21] and Uncalibrated PS (UPS) [1,17,37,38] achieve high precision under idealized lighting models (e.g., directional illumination), but their reliance on strong lighting assumptions limits real-world applicability. To address this limitation,

* Xiaoyan Liu, Yujie Xing and Qian Chen are equal corresponding authors.

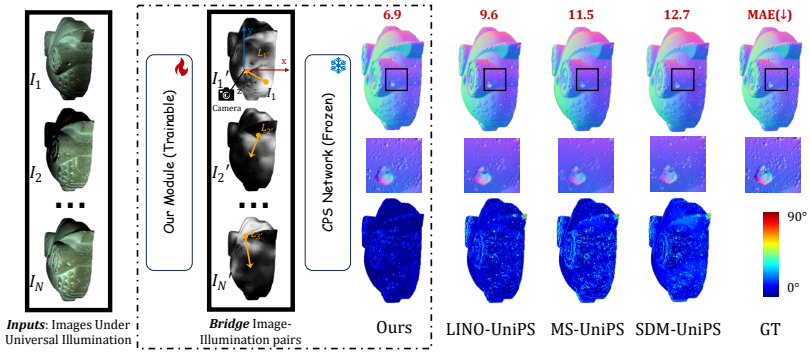


Fig. 1: Overview of Bridge-UniPS. We transform unconstrained images into CPS-compatible image-illumination pairs. The adapter leverages inherent physical priors in pretrained CPS models to guide the illumination-consistent transformation.

Universal PS (UniPS) [18, 19] has emerged as a practical paradigm that aims to recover high-fidelity normals directly from images captured under arbitrary, unknown illumination. By removing controlled acquisition requirements, UniPS enables scalable 3D capture in unconstrained environments.

Most existing UniPS methods [11, 28] adopt end-to-end learning formulations that implicitly infer illumination while predicting surface normals. These approaches typically rely on global illumination contexts or latent tokens to guide normal estimation.

Despite promising benchmark results, such end-to-end designs exhibit two inherent limitations. Firstly, the absence of explicit, physically grounded intermediate representations forces illumination and geometry to be learned in a **highly coupled manner**, making models sensitive to distribution shifts and prone to geometric ambiguities in real-world scenarios. Secondly, representing illumination through global statistical summaries often **suppresses local photometric variations** that are essential for recovering fine-scale geometry.

Motivated by the limitations of end-to-end paradigms, we propose **Bridge-UniPS**, which introduces a lighting adapter to transform raw observations into image–illumination pairs that are directly compatible with CPS models. This establishes a **physics-aligned intermediate representation**, allowing the adapter to be guided by the physical priors implicitly encoded in CPS networks and thereby improving the accuracy of surface normal recovery, as illustrated in Fig. 1.

To preserve high-frequency geometry, we design a **Dynamic Neural Query** module. Unlike prior approaches that aggregate features globally, this module uses predefined target light directions as queries to adaptively retrieve physically informative fragments from raw observations. This selective retrieval mechanism filters out redundant noise and forms a more robust feature representation.

Crucially, the adapter is optimized in a task-driven manner without direct intermediate supervision. During training, the final normal-estimation objective

is supervised by ground-truth normals, while no ground-truth Bridge Images or intermediate illumination annotations are provided. The adapter is optimized through gradients propagated through a frozen pre-trained CPS network, which constrains the generated image–illumination pairs to remain compatible with CPS inference. Despite the absence of direct supervision on intermediate representations, the adapter learns to generate light-conditioned and CPS-compatible Bridge Images that preserve image–illumination consistency, suggesting that modern CPS models can provide useful priors to guide this transformation.

Our main contributions are summarized as follows:

- We propose a bridging paradigm for universal photometric stereo, which maps unconstrained observations into a CPS-compatible intermediate representation, allowing high-precision CPS models to be applied in unconstrained settings.
- We introduce a Dynamic Neural Query module that selectively retrieves informative features based on target light directions, enabling more robust representations of both global and local information.
- We develop a task-driven lighting transformation framework through a frozen CPS network, in which the lighting adapter is optimized without direct supervision on intermediate Bridge Images or illumination annotations, enabling the learned Bridge Image–Illumination pairs to serve as CPS-compatible intermediate representations.

2 Related Works

2.1 Calibrated Photometric Stereo

Calibrated photometric stereo (CPS) recovers surface normals from multiple images captured under known illumination conditions, typically assuming directional or point light sources with calibrated parameters [43]. Early extensions of Woodham's paradigm addressed increasingly complex reflectance behaviors, including non-Lambertian materials and specularities, through robust estimation [44], reflectance modeling [20], and example-based priors [15]. These approaches [13, 40] explicitly enforce physical constraints at each surface point, enabling accurate normal estimation in controlled environments.

With the emergence of deep learning [12,27], modern CPS methods [4,14,45] further improve the performance on complex BRDFs [32,39,46] while still relying on calibrated lighting, thereby preserving physical interpretability. The requirement for precise lighting calibration largely confines CPS methods to laboratory settings, limiting their applicability in uncontrolled real-world scenarios.

2.2 Uncalibrated Photometric Stereo

To relax the reliance on calibrated lighting, uncalibrated photometric stereo (UPS) [2, 7, 9] was introduced. While UPS methods extend CPS to more complex illumination scenarios [3, 5, 10, 26, 30, 31], they remain limited under unconstrained lighting, primarily due to the difficulty of accurately modeling arbitrary and intricate light interactions. In this sense, UPS serves as a transitional

paradigm between the physically constrained CPS and the fully data-driven universal photometric stereo (UniPS), motivating methods that can generalize to unknown illumination environments.

2.3 Universal Photometric Stereo

Universal photometric stereo (UniPS) aims to recover surface normals under arbitrary and unknown illumination without predefined light models [18]. Early encoder–decoder based approaches [18, 19] aggregate multi-illumination observations into global feature representations for normal prediction. More recent transformer-based methods, such as LINO-UniPS [28], attempt to explicitly model illumination by introducing light register tokens and interleaved attention mechanisms [6, 42] and achieve superior results.

Despite these advances, most UniPS methods lack explicit intermediate physical constraints, causing illumination and geometry cues to remain entangled in the learned representations, which can limit robustness and generalization under complex real-world lighting.

3 Method

3.1 Problem Definition and Overview

We consider the **Universal Photometric Stereo (UniPS)** problem, which aims to recover a high-quality surface normal map $\mathbf{N} \in \mathbb{R}^{3\times H\times W}$ from a sequence of images $\mathbf{I} \in \mathbb{R}^{N\times 3\times H\times W}$ captured under unknown and unconstrained illumination.

As illustrated in Fig. 2, our framework consists of two stages: a *trainable task-driven lighting transformation* stage and a *frozen calibrated photometric stereo (CPS)* stage.

In the first stage, a learnable **lighting adapter** converts the unconstrained image sequence into a set of **Bridge Image–Illumination pairs**, which serve as an explicit CPS-compatible intermediate representation. This transformation is implemented by our **Dynamic Neural Query** module, which selectively extracts and aggregates informative features from the input sequence conditioned on target illumination directions, and decodes them into light-conditioned Bridge Images paired with corresponding illumination vectors.

In the second stage, these Bridge Image–Illumination pairs are directly fed into a frozen CPS network to reconstruct the target surface normals.

3.2 Task-driven Lighting Transformation Stage

Input: Image sequence $\mathbf{I} \in \mathbb{R}^{N_{\text{in}}\times 3\times H\times W}$.

Outputs: Bridge Images $\mathbf{B} \in \mathbb{R}^{N_{\text{out}}\times 3\times H\times W}$ with corresponding **illumination directions** $\mathbf{L} \in \mathbb{R}^{N_{\text{out}}\times 3}$.

This stage trains a lighting adapter without direct supervision on the output Bridge Images or their paired illumination directions. Instead, it is optimized through the final normal-estimation objective, with gradients propagated through a frozen downstream CPS network. This encourages the generated Bridge Image–Illumination pairs to remain compatible with CPS inference.

Feature Extraction Given the input image sequence $\mathbf{I}$, we extract convolutional features using a ConvNeXt [33] backbone network, producing a feature tensor $\mathbf{F} \in \mathbb{R}^{N_{\text{in}} \times C \times L}$. Based on these features, we further derive three types of representations: **Light Feat** $\mathbf{F}_l \in \mathbb{R}^{N_{\text{light}} \times C \times L}$, **Pixel Feat** $\mathbf{F}_p \in \mathbb{R}^{N_{\text{light}} \times C \times L}$, and a **Global Conf** $\mathbf{G} \in \mathbb{R}^{N_{\text{light}} \times N_{\text{in}}}$, where $C = 1024$ and $L = HW/1024$.
Light Feat. To bridge unconstrained capture with CPS, we define a fixed set of N_{light} target directions. Based on the observation that $> 95\%$ of lighting directions in common PS datasets lie within the spherical cap ($z > 0.6$), we perform N_{light} uniform area-preserving sampling in this region. These directions are mapped via an MLP into the channel space C to produce the **Light Feat** $\mathbf{F}_l \in \mathbb{R}^{N_{\text{light}} \times C \times L}$.
Other Features. To capture cross-image dependencies, we apply a self-attention operation on $\mathbf{F}$ and average it along the N_{in} dimension. The resulting tensor is expanded to size N_{light} to form the **Pixel Feat** $\mathbf{F}_p \in \mathbb{R}^{N_{\text{light}} \times C \times L}$, providing sequence-level pixel information. Additionally, to assess image reliability, we map the C and L dimensions of $\mathbf{F}$ to scalars using MLPs, generating a global quality score per image. These scores are expanded into the **Global Conf** matrix $\mathbf{G} \in \mathbb{R}^{N_{\text{light}} \times N_{\text{in}}}$, reflecting the MLP-based assessment of each input image's reliability under different target illuminations.

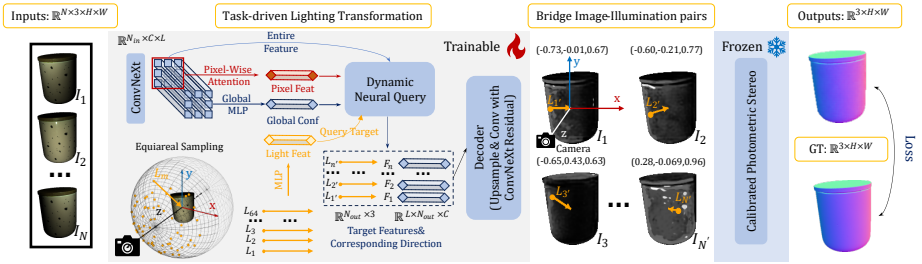
Dynamic Neural Query Given illumination embeddings $\mathbf{F}_l \in \mathbb{R}^{N_{\text{light}} \times C}$ and image features $\mathbf{F} \in \mathbb{R}^{N_{\text{in}} \times C \times L}$, our goal is to retrieve, for each target illumination, the most relevant feature evidence from the unconstrained image sequence.

Light–Image Compatibility. We first estimate how compatible each input image is with each target illumination. To this end, we derive a refined key representation $\mathbf{K} \in \mathbb{R}^{N_{\text{in}} \times C \times L}$ from $\mathbf{F}$ via a channel-wise MLP and a spatial convolution, which enhances pixel-level discriminability. We then compute its spatial mean to obtain a global key $\bar{\mathbf{K}} \in \mathbb{R}^{N_{\text{in}} \times C}$. Together with the illumination embedding $\mathbf{F}_l$, we compute a similarity matrix

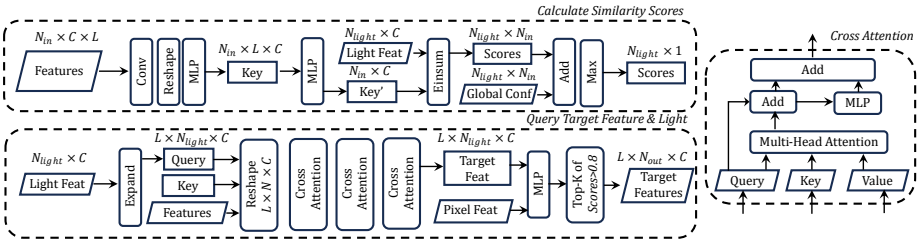
$$\mathbf{S} = \mathbf{F}_l \bar{\mathbf{K}}^\top + \log(\mathbf{G}), \tag{1}$$

where $\mathbf{G} \in \mathbb{R}^{N_{\text{light}} \times N_{\text{in}}}$ denotes the Global Conf matrix that encodes the overall reliability of each input image. The score $S_{j,i}$ reflects how suitable image i is for explaining observations under illumination j.

Illumination-conditioned Retrieval. To retrieve pixel-level evidence, we perform cross-attention conditioned on each target illumination. The illumination embedding $\mathbf{F}_l$ is expanded along the spatial dimension to form the query tensor



(a) Overall pipeline of Bridge-UniPS.



(b) Architecture of the Dynamic Neural Query module for lighting-conditioned feature selection.

Fig. 2: Overview of the proposed Bridge-UniPS framework. (a) The two-stage pipeline that bridges universal photometric stereo to calibrated photometric stereo via a task-driven lighting adapter. (b) Detailed structure of the Dynamic Neural Query module used in the lighting transformation stage.

$\mathbf{Q} \in \mathbb{R}^{L\times N_{\text{light}}\times C}$, while the refined key $\mathbf{K}$ and the original ConvNeXt features $\mathbf{F}$ are reshaped into $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{L\times N_{\text{in}}\times C}$ and used as the key and value, respectively. Cross-attention is then computed as $\mathbf{T} = \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, yielding illumination-specific target features $\mathbf{T} \in \mathbb{R}^{L\times N_{\text{light}}\times C}$.

Score-guided Feature Selection. Not all input observations are consistent with a given illumination. We therefore use the similarity scores $\mathbf{S}$ to filter unreliable features. For each illumination, we select candidates whose scores exceed a threshold of 0.8 and then keep the top-ranked ones, subject to a minimum and maximum number constraint to ensure stable downstream processing. This produces a binary selection mask that suppresses incompatible image contributions and retains the most relevant feature evidence.

The filtered target features are finally passed to the decoder to generate the corresponding bridge image–illumination pairs.

Decoder with Attention-guided Skip Connections The decoder mirrors the ConvNeXt encoder in a symmetric fashion, where each downsampling stage is inverted by an upsampling block composed of bilinear upsampling followed by convolution.

From the cross-attention module, we obtain illumination-conditioned attention weights $\mathbf{A} \in \mathbb{R}^{L \times N_{\text{out}} \times N_{\text{in}}}$, which quantify, at each spatial location, how strongly each input image contributes to a selected target illumination.

Since the values in attention are directly taken from ConvNeXt feature maps, we can reuse them to construct illumination-aware skip connections. For each decoder level s, the attention weights are resized to match the spatial resolution $L_s = H_s W_s$ of the corresponding skip features, yielding $\mathbf{A}^{(s)} \in \mathbb{R}^{B \times N_{\text{out}} \times N_{\text{in}} \times L_s}$. Let the skip features be $\mathbf{F}^{(s)} \in \mathbb{R}^{B \times N_{\text{in}} \times C \times L_s}$. We aggregate them by

$$\mathbf{F}^{(s)}_{\text{skip}} = \sum_{i=1}^{N_{\text{in}}} \mathbf{A}^{(s)}_{:,:,i,:} \odot \mathbf{F}^{(s)}_{:,i,:,:}, \tag{2}$$

and feed the resulting $B \times N_{\text{out}} \times C \times L_s$ features into the decoder.

This enables each bridge image to be reconstructed from multi-scale ConvNeXt features that are dynamically selected and spatially weighted according to the target illumination.

3.3 Frozen Calibrated Photometric Stereo Stage

Input: Bridge Images $\mathbf{B} \in \mathbb{R}^{N_{\text{out}} \times 3 \times H \times W}$ with corresponding **illumination directions** $\mathbf{L} \in \mathbb{R}^{N_{\text{out}} \times 3}$. **Outputs: Surface Normal Maps** $\mathbf{N} \in \mathbb{R}^{3 \times H \times W}$. Let $\mathcal{P}$ denote a calibrated photometric stereo (CPS) network pretrained under standard point-light conditions. In Bridge-UniPS, $\mathcal{P}$ serves as the final normal estimator for the Universal PS task and is kept entirely frozen.

Given the predicted bridge images $\mathbf{I}_b$ together with their associated illumination directions $\mathbf{L}_b$, the CPS module outputs surface normals

$$\mathbf{N} = \mathcal{P}(\mathbf{I}_b, \mathbf{L}_b) \in \mathbb{R}^{3 \times H \times W}, \tag{3}$$

which directly correspond to the target of photometric stereo.

The training objective is defined only on these normals:

$$\mathcal{L} = \gamma_1 \, \mathcal{L}_{\cos}(\mathbf{N}, \mathbf{N}_{gt}) + \gamma_2 \, \|\mathbf{N} - \mathbf{N}_{gt}\|_1. \tag{4}$$

No direct supervision is applied to the intermediate bridge images or illumination estimates.

As a result, the lighting adaptation stage is optimized solely through gradients back-propagated from the frozen CPS module. This treats $\mathcal{P}$ not only as the final reconstruction operator, but also as a physics-aligned constraint that guides the learning of relighting and light–geometry decoupling.

4 Results

4.1 Experimental Setup

Training Details Our model is primarily trained on the training split of the Uni-MS-PS dataset [11]. To preserve performance under classical point-light conditions, we additionally include approximately 10,000 samples from PS-Sculpture and PS-Blobby [4], accounting for about 10% of the total training set.

Table 1: Quantitative (MAE °) on the DiLiGenT benchmark. The best results are highlighted in **bold**. We achieve the best average error across all objects.

Method	Task	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Ave
LL22a [29]	Calibrated	2.4	3.6	8.0	4.9	4.7	6.7	14.9	6.0	5.0	8.8	6.5
JS22 [25]		*2.9*	*5.5*	*7.1*	*4.7*	*6.0*	*7.5*	*12.3*	*6.0*	*6.4*	*9.9*	*6.8*
CH19 [16]	Uncalibrated	2.8	6.9	9.0	8.1	8.5	11.9	17.4	8.1	7.5	14.9	9.5
KK21 [26]		3.8	6.0	13.1	7.9	10.9	11.9	25.5	8.8	10.2	18.2	11.6
LL22b [30]		**1.2**	3.8	9.3	4.7	5.5	7.1	14.6	6.7	6.5	10.5	7.1
JS23 [24]		2.2	5.6	6.7	4.3	6.1	6.8	12.0	5.5	6.4	9.7	6.6
IK22 [18]	Universal	4.9	9.1	19.4	13.0	11.6	24.2	25.2	10.8	9.9	18.8	14.7
IK23 [19]		1.5	3.6	7.5	5.4	4.5	8.5	10.2	4.7	4.1	8.2	5.8
LC25 [28]		1.8	2.6	6.2	3.4	4.3	**5.2**	8.6	**4.1**	4.6	6.7	4.7
Ours	Universal	2.0	**2.5**	**5.2**	**3.0**	**3.2**	6.2	**6.5**	4.7	**4.0**	**5.2**	**4.3**
Ours (K=16)		2.0	2.6	5.3	3.1	3.2	6.5	6.7	4.7	4.0	5.5	4.5
Ours (K=4)		2.1	4.1	7.2	4.6	5.1	8.8	9.1	6.5	5.4	7.6	6.0

The network is optimized using the AdamW optimizer [34] with an initial learning rate of 1×10^{-5}, together with a step decay schedule that multiplies the rate by 0.8 every 20 epochs. The batch size is set to 12. For each batch, the number of input images is randomly sampled from 6 to 24, which improves robustness to varying numbers of observations. The model is trained for 140 epochs. All experiments are conducted on four NVIDIA A100 GPUs, and the full training process takes approximately 9 days.

The training objective is defined on the surface normals predicted by the frozen CPS network, using the loss specified in Eq. 4.

Key Hyperparameters For feature extraction, we use the ConvNeXt Base architecture from `torchvision.models.convnext.ConvNeXt` [36], which produces feature maps with 1024 channels. The decoder is constructed by mirroring this backbone, where each convolutional block is replaced with an upsampling layer followed by a convolution. The number of equiareal illumination samples is set to $N_{\text{light}} = 64$, uniformly distributed over the region of the unit sphere with $z > 0.6$, as described earlier. We use the pretrained Norm-PSN [25] as the frozen CPS network.

4.2 Constrained Illumination Benchmarks

DiLiGenT We first evaluate our method on the DiLiGenT [41] benchmark. This benchmark contains 10 datasets, covering objects with diverse surface geometries and reflectance properties. Each dataset provides 96 images of resolution 612×512 in 16-bit HDR format, captured under known single directional lighting.

Among the competing methods, the selected calibrated and uncalibrated approaches are specifically designed for single directional lighting. The distinction

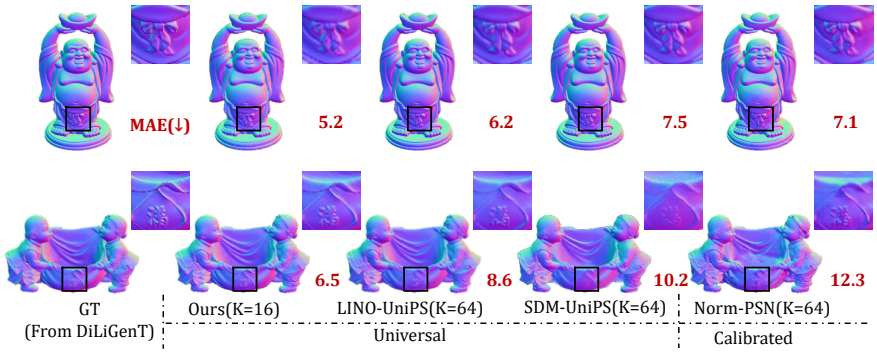


Fig. 3: Qualitative comparison on the DiLiGenT benchmark. Our method achieves state-of-the-art performance while using only 16 input images, significantly outperforming existing methods that use the full set of 96 images.

between calibrated and uncalibrated methods is based solely on whether ground-truth lighting directions are required as input, regardless of whether lighting-related variables are explicitly or implicitly estimated during their internal processing. Universal methods, in contrast, do not require lighting calibration and are capable of handling images captured under unconstrained illumination.

To analyze the influence of the number of input images, we report results of our method under varying input counts. All other methods are evaluated using the full set of 96 images.

The quantitative results in terms of mean angular error are reported in Table 1. As shown in Table 1, our method achieves state-of-the-art performance both on average and across most datasets, outperforming all existing approaches.

Benefiting from the lighting adapter that remaps the input observations into CPS-compatible representations, our method maintains state-of-the-art accuracy even when using only 16 input images. Further reducing the number of input images has only a limited impact on performance, demonstrating the robustness of the proposed framework.

Qualitative comparisons are presented in Fig. 3, where our method uses only 16 input images. On two of the more challenging datasets in this benchmark (*Harvest* and *Buddha*), existing methods often exhibit degraded reconstruction quality in fine-scale details. In contrast, our approach reconstructs sharper and more geometrically faithful normal maps with fewer input images, producing results that are closer to the ground truth and highlighting the effectiveness of the proposed bridging strategy.

Illumination Statistics Stabilization Analysis To further investigate the mechanism behind the performance improvement, we analyze how the proposed lighting adapter affects illumination contrast statistics in the DiLiGenT benchmark. For each image, we compute the standard deviation of masked intensity (i.e., contrast) within the valid object region. We then summarize three dataset-

Table 2: Cross-object illumination contrast statistics on DiLiGenT. More consistent values in the same row indicate improved cross-object illumination stability.

Inputs	Stat	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading
Original	MEAN	0.04	0.03	0.04	0.03	0.04	0.05	0.05	0.01	0.01	0.08
	STD	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.01	0.01	0.03
	MED	0.04	0.03	0.03	0.03	0.03	0.04	0.05	0.01	0.01	0.08
Bridge	MEAN	0.30	0.33	0.35	0.31	0.31	0.37	0.31	0.34	0.35	0.33
	STD	0.03	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.03
	MED	0.30	0.34	0.35	0.32	0.32	0.37	0.31	0.34	0.35	0.32

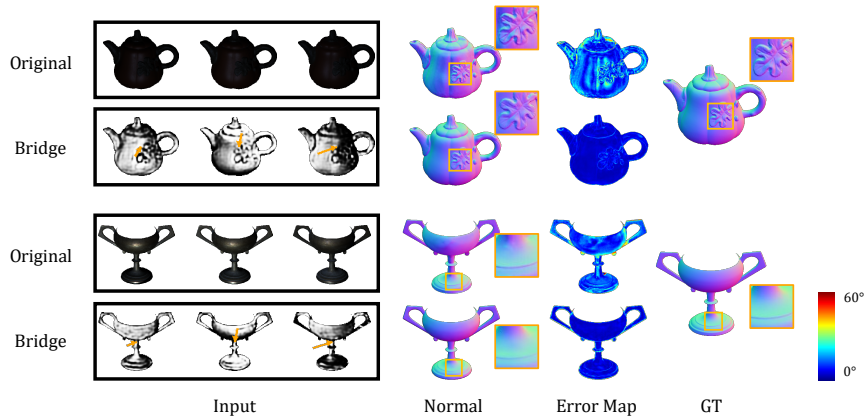


Fig. 4: Qualitative comparison between original DiLiGenT inputs and bridge images. Our lighting adapter produces visually realistic images with stabilized illumination statistics, leading to more detailed and accurate normal reconstruction.

level statistics for each object: the mean, standard deviation, and median across all lighting conditions.

Table 2 reports the contrast statistics for all objects in DiLiGenT. The original benchmark exhibits substantial cross-object variation due to diverse reflectance properties, particularly for highly specular objects such as *Goblet*. In contrast, the bridge images generated by our lighting adapter demonstrate significantly more consistent contrast distributions across objects. This indicates that the adapter implicitly stabilizes illumination statistics and reduces reflectance-induced discrepancies.

To isolate this effect, we feed both the original inputs and the bridge images into the same pretrained and frozen Norm-PSN model. The network architecture and weights remain unchanged. As shown in Table 1, replacing the raw inputs (JS22 [25]) with bridge images (ours) leads to clear improvements in normal estimation accuracy, confirming that the adapter enhances compatibility with CPS rather than relying on architectural modifications. Qualitative comparisons are shown in Fig. 4. Objects such as *Goblet* (metallic, non-Lambertian) and *Pot2* (approximately Lambertian) exhibit drastically different contrast lev-

Table 3: Quantitative (MAE °) comparison on the LUCES benchmark. Our method achieves state-of-the-art performance on most of datasets and the overall average.

Method	Ball	Bell	Bowl	Buddha	Bunny	Cup	Die	Hippo	House	Jar	Owl	Queen	Squirrel	Tool	Avg.
IK22 [18]	11.01	24.12	23.84	27.90	23.51	28.64	16.24	21.41	35.93	14.53	32.87	28.36	25.36	19.03	23.77
IK23 [19]	13.30	12.76	8.44	18.58	8.53	19.67	7.25	8.86	26.07	8.30	12.67	15.97	16.01	12.54	13.50
HQ24 [11]	10.20	10.52	6.98	12.83	9.60	13.68	6.19	8.33	25.29	6.30	11.47	12.45	11.36	11.79	11.21
LC25 [28]	10.16	8.78	6.96	12.67	**6.09**	8.15	6.16	5.99	22.91	6.24	9.58	9.84	10.25	8.25	9.43
Ours	**9.25**	**4.78**	**3.97**	**12.06**	6.91	**4.92**	**4.38**	**4.90**	**16.08**	**3.83**	**6.92**	**8.68**	**7.43**	**5.36**	**7.11**

els in the original benchmark. After adaptation, the illumination appears more stable and point-light-like, resulting in more detailed and geometrically faithful normal reconstructions.

4.3 Unconstrained Illumination Benchmarks

LUCES (Near-Field Lighting) Unlike controlled far-field illumination, which is relatively stable and distance-invariant, near-field lighting exhibits strong source–surface coupling effects, leading to highly sensitive and nonlinear intensity variations. Such characteristics make near-field illumination a representative and challenging case of unconstrained lighting.

We first evaluate our method on the LUCES [35] benchmark, which is captured under near-field lighting conditions. This benchmark contains 14 datasets, each providing 52 large-scale images with a resolution of 2048×1536, along with corresponding object masks and calibrated light source information.

Since near-field illumination violates the assumptions adopted by most CPS and UPS methods, we compare only with existing universal methods. All methods are evaluated using as many input images as possible.

The quantitative results are reported in Table 3. As shown in the table, our method achieves state-of-the-art performance on 13 out of 14 datasets as well as on the overall average, outperforming all existing universal approaches. These results demonstrate the effectiveness of the proposed pipeline in transforming unconstrained observations into controllable image–illumination pairs, enabling robust normal estimation under challenging lighting conditions.

The *Bell* and *Cup* datasets consist of metallic objects exhibiting strong non-Lambertian reflectance. Compared to competing methods, our approach yields significantly improved performance on these datasets, further highlighting the robustness of the proposed bridging strategy in handling complex reflectance behaviors.

Qualitative comparisons are presented in Fig. 5. We select the structurally and texturally complex HOUSE and SQUIRREL datasets for visualization, together with their corresponding error maps. As shown in the figure, our method reconstructs more detailed geometric structures and produces surface normals that are closer to the ground truth, especially in fine-scale regions, demonstrating its superior performance under near-field and highly unconstrained illumination.

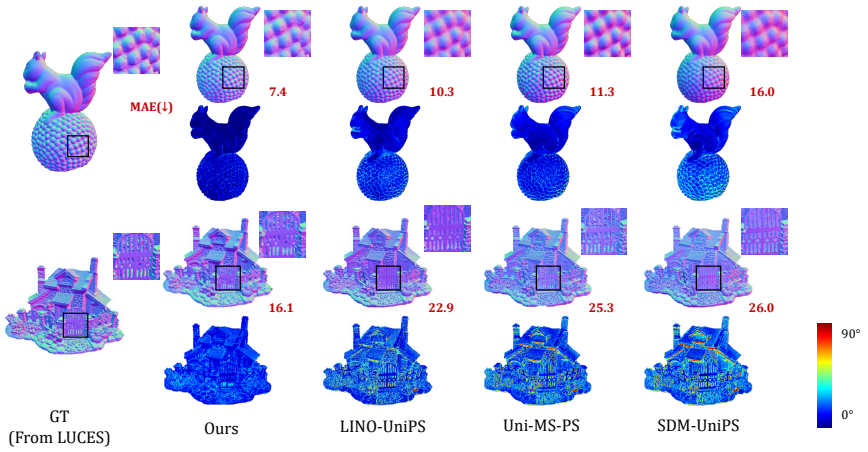


Fig. 5: Qualitative comparison on the LUCES benchmark. Our method produces more geometric details and lower angular errors.

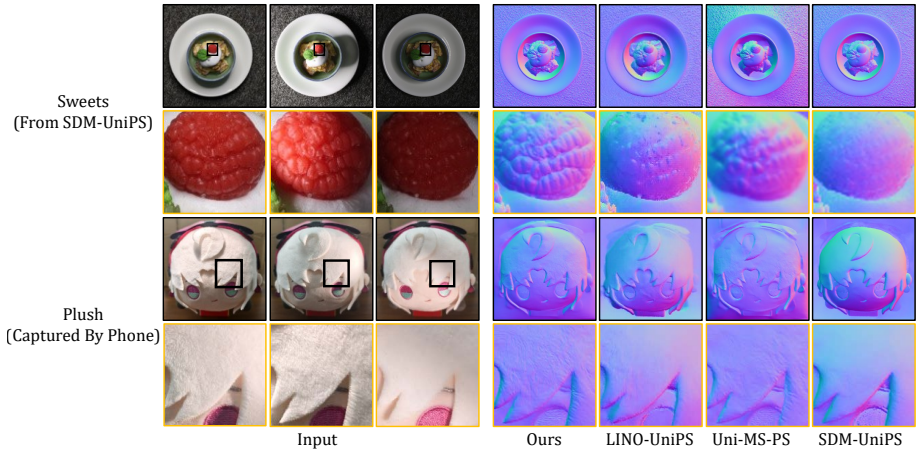


Fig. 6: Qualitative results on real-world scenes. Our method reconstructs detailed and stable surface normals even under relatively weak lighting variation.

Evaluation on Real-World Scenes To further evaluate the generalization ability of our method under real-world conditions, we conduct experiments on two practical scenes: the *Sweets* scene provided by SDM-UniPS and a self-captured *Plush* scene. Unlike benchmark datasets with calibrated setups, these real scenes contain uncontrolled environmental light, highlighting practical applicability. Qualitative results are shown in Fig. 6.

The two scenes exhibit distinct illumination characteristics. The Sweets scene is captured under relatively weak ambient light with strong directional illumination, whereas the Plush scene is intentionally recorded under stronger ambient

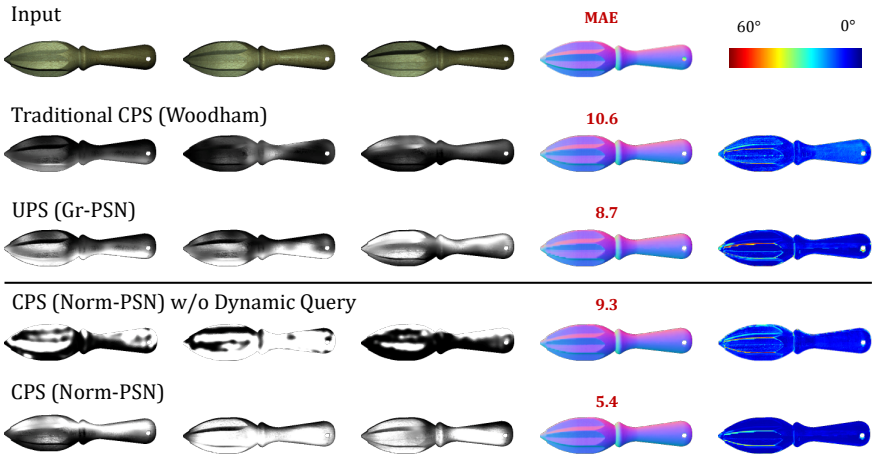


Fig. 7: Ablation study on downstream photometric stereo paradigms and lighting adapter components.

lighting with comparatively weaker light source variation. This setting allows us to assess robustness under reduced lighting contrast and less pronounced shading cues.

As shown in Fig. 6, our method consistently reconstructs surface normals with richer geometric details in both scenarios. Notably, in the Plush scene where lighting variation is relatively subtle, our approach still produces stable and coherent normals even on approximately planar regions. These results demonstrate the robustness of the proposed pipeline across diverse real-world illumination conditions.

4.4 Ablation Study

Impact of Downstream Photometric Stereo Paradigms We study how different downstream photometric stereo (PS) paradigms influence both the generated bridge images and the final normal estimation accuracy. Three paradigms are considered: traditional CPS, learning-based UPS, and learning-based CPS. Table 4 reports results using representative methods from each paradigm, including Woodham (1980) [43], SCPS-NIR [30], GR-PSN [24], CNN-PS [16], and Norm-PSN [25]. All models are trained under identical configurations.

As shown in Fig. 7, all paradigms generate bridge images approximating parallel illumination rather than the distant point-light setup used in acquisition, consistent with the PS assumption. For visualization clarity, the figure displays three representative methods (one per paradigm), while Table 4 provides quantitative results for all evaluated models.

Interestingly, traditional CPS produces darker bridge images with stronger non-Lambertian artifacts, reflecting its strict Lambertian constraint that forces

Table 4: Ablation study of downstream PS paradigms, lighting adapter components, and additional control settings. MAE (in degrees, ↓) is reported.

Setting	Paradigm / Variant	DiLiGenT	LUCES
Downstream PS Methods			
Woodham (1980) [43]	Traditional CPS	9.2	14.1
SCPS-NIR [30]	Learning-based UPS	6.5	12.3
GR-PSN [24]	Learning-based UPS	5.8	10.8
CNN-PS [16]	Learning-based CPS	5.4	9.7
Norm-PSN [25] (*Full model*)	Learning-based CPS	**4.3**	**7.1**
Module Design			
w/o global conf	Learning-based CPS	4.7	8.1
w/o pixel feat	Learning-based CPS	4.8	7.9
w/o top-K selection	Learning-based CPS	5.6	9.4
w/o dynamic query	Learning-based CPS	6.3	12.5
Additional Control Ablations			
w/o point-light data	Training data	4.4	7.1
learnable latent tokens	Target-light conditioning	4.7	8.4
unfrozen CPS	CPS optimization	5.4	7.9
shuffled bridge pairing	Image–light consistency	44.5	43.8

the adapter to approximate arbitrary materials with Lambertian-consistent representations. In contrast, learning-based CPS and UPS produce bridge images lying between Lambertian and non-Lambertian reflectance, likely due to normalization layers [22] and training on mixed-material datasets that implicitly relax reflectance assumptions.

Quantitative results in Table 4 show that the overall MAE trend correlates with PS paradigms: learning-based methods significantly outperform traditional CPS, and learning-based CPS consistently surpasses learning-based UPS, indicating that calibrated supervision provides stronger structural guidance for learning consistent image–illumination pairs.

Effect of Lighting Adapter Components We further analyze the contribution of each component in the lighting adapter. For *w/o global confidence* and *w/o pixel features*, the corresponding feature branches are removed from the fusion pipeline by directly modifying the input channel dimensions of the downstream MLP. For *w/o top-K selection*, all generated bridge images are used without confidence-based filtering. Finally, *w/o dynamic query* disables the entire dynamic querying mechanism, reducing it to a simple stack of attention layers without adaptive selection or structured supervision.

Quantitative results in Table 4 show consistent performance degradation when any component is removed. Among them, removing the dynamic query causes the most significant drop on both benchmarks, highlighting its central role in stabilizing image–illumination pair learning.

As illustrated in Fig. 7, when the dynamic query is disabled, the generated bridge images lose illumination stability and no longer exhibit the expected parallel-light characteristics. The lighting becomes visually inconsistent and less physically plausible, which leads to degraded normal reconstruction. This con-

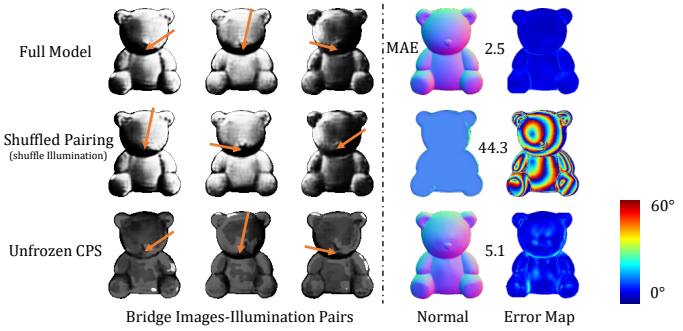


Fig. 8: Additional control ablations on shuffled bridge pairing and unfrozen CPS.

firms that the dynamic query is essential for enforcing structured illumination adaptation and ensuring reliable supervision from the downstream PS model.

Additional Control Ablations We add several control ablations to examine whether the improvement can be explained by alternative factors such as additional point-light training data, generic query-based routing, adapter–CPS co-adaptation, or bridge-image generation alone. First, removing the additional point-light data only causes a modest change, suggesting that the performance gain is not mainly due to extra exposure to classical PS samples. Second, replacing target-light directions with learnable latent tokens degrades the results, especially on LUCES, indicating that explicit target-light parameterization provides useful inductive bias beyond generic latent routing. Third, unfreezing the CPS backbone leads to worse performance, suggesting that keeping the CPS model frozen helps prevent excessive adapter–CPS co-adaptation. Finally, shuffling the Bridge Image–Illumination pairings causes catastrophic degradation, confirming that the generated Bridge Images are not merely favorable surrogate images; their consistency with the paired target lights is essential for CPS-compatible inference, as shown in Fig. 8.

5 Conclusion

This paper presents Bridge-UniPS, a novel framework bridging unconstrained capture and calibrated physical models via a task-driven lighting adapter. By mapping complex observations into a physics-aligned intermediate space, it achieves state-of-the-art performance across diverse benchmarks, maintaining high geometric fidelity and robust generalization in unconstrained illumination.

Despite these improvements, weak photometric variation—such as scenes dominated by ambient illumination with minimal active lighting—remains challenging. Future work will focus on improving the sensitivity of the lighting adapter to subtle illumination cues, enabling more stable reconstruction under minimal lighting diversity.

References

1. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. International journal of computer vision **35**(1), 33–44 (1999)
2. Chandraker, M.K., Kahl, F., Kriegman, D.J.: Reflections on the generalized bas-relief ambiguity. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 788–795. IEEE (2005)
3. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Self-calibrating deep photometric stereo networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8739–8747 (2019)
4. Chen, G., Han, K., Wong, K.Y.K.: Ps-fcn: A flexible learning framework for photometric stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–18 (2018)
5. Chen, G., Waechter, M., Shi, B., Wong, K.Y.K., Matsushita, Y.: What is learned in deep uncalibrated photometric stereo? In: European Conference on Computer Vision. pp. 745–762. Springer (2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Drbohlav, O., Šára, R.: Specularities reduce ambiguity of uncalibrated photometric stereo. In: European Conference on Computer Vision. pp. 46–60. Springer (2002)
8. Enomoto, K., Waechter, M., Kutulakos, K.N., Matsushita, Y.: Photometric stereo via discrete hypothesis-and-test search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2311–2319 (2020)
9. Georghiades: Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In: Proceedings Ninth IEEE International Conference on Computer Vision. pp. 816–823. Ieee (2003)
10. Haefner, B., Ye, Z., Gao, M., Wu, T., Quéau, Y., Cremers, D.: Variational uncalibrated photometric stereo under general lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8539–8548 (2019)
11. Hardy, C., Quéau, Y., Tschumperlé, D.: Uni ms-ps: A multi-scale encoder-decoder transformer for universal photometric stereo. Computer Vision and Image Understanding **248**, 104093 (2024)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hertzmann, A., Seitz, S.M.: Example-based photometric stereo: Shape reconstruction with general, varying brdfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1254–1264 (2005)
14. Honzátko, D., Türetken, E., Fua, P., Dunbar, L.A.: Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo. In: 2021 International Conference on 3D Vision (3DV). pp. 394–402. IEEE (2021)
15. Hui, Z., Sankaranarayanan, A.C.: Shape and spatially-varying reflectance estimation from virtual exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(10), 2060–2073 (2016)
16. Ikehata, S.: Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–18 (2018)

17. Ikehata, S.: Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism. arXiv preprint arXiv:2211.11386 (2022)
18. Ikehata, S.: Universal photometric stereo network using global lighting contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12591–12600 (2022)
19. Ikehata, S.: Scalable, detailed and mask-free universal photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13198–13207 (2023)
20. Ikehata, S., Aizawa, K.: Photometric stereo using constrained bivariate regression for general isotropic surfaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2179–2186 (2014)
21. Ikehata, S., Wipf, D., Matsushita, Y., Aizawa, K.: Robust photometric stereo using sparse regression. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 318–325. IEEE (2012)
22. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
23. Ju, Y., Lam, K.M., Xie, W., Zhou, H., Dong, J., Shi, B.: Deep learning methods for calibrated photometric stereo and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
24. Ju, Y., Shi, B., Chen, Y., Zhou, H., Dong, J., Lam, K.M.: Gr-psn: Learning to estimate surface normal and reconstruct photometric stereo images. IEEE Transactions on Visualization and Computer Graphics (2023)
25. Ju, Y., Shi, B., Jian, M., Qi, L., Dong, J., Lam, K.M.: Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention. International Journal of Computer Vision **130**(12), 3014–3034 (2022)
26. Kaya, B., Kumar, S., Oliveira, C., Ferrari, V., Van Gool, L.: Uncalibrated neural inverse rendering for photometric stereo of general surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3804–3814 (2021)
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
28. Li, H., Chen, H., Ye, C., Chen, Z., Li, B., Xu, S., Guo, X., Liu, X., Wang, Y., Zhang, B., et al.: Light of normals: Unified feature representation for universal photometric stereo. arXiv preprint arXiv:2506.18882 (2025)
29. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16221–16230 (2022)
30. Li, J., Li, H.: Self-calibrating photometric stereo by neural inverse rendering. In: European Conference on Computer Vision. pp. 166–183. Springer (2022)
31. Li, Z., Zheng, Q., Shi, B., Pan, G., Jiang, X.: Dani-net: Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8381–8391 (2023)
32. Liu, H., Yan, Y., Song, K., Yu, H.: Sps-net: Self-attention photometric stereo network. IEEE Transactions on Instrumentation and Measurement **70**, 1–13 (2020)
33. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)

34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
35. Mecca, R., Logothetis, F., Budvytis, I., Cipolla, R.: Luces: A dataset for near-field point light source photometric stereo. arXiv preprint arXiv:2104.13135 (2021)
36. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
37. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 283–291 (2018)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
39. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 501–509 (2017)
40. Shi, B., Tan, P., Matsushita, Y., Ikeuchi, K.: Bi-polynomial modeling of low-frequency reflectances. IEEE transactions on pattern analysis and machine intelligence **36**(6), 1078–1091 (2013)
41. Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3707–3716 (2016)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
43. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical engineering **19**(1), 139–144 (1980)
44. Wu, L., Ganesh, A., Shi, B., Matsushita, Y., Wang, Y., Ma, Y.: Robust photometric stereo via low-rank matrix completion and recovery. In: Asian conference on computer vision. pp. 703–717. Springer (2010)
45. Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B.: Gps-net: Graph-based photometric stereo network. Advances in Neural Information Processing Systems **33**, 10306–10316 (2020)
46. Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L.Y., Kot, A.C.: Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8549–8558 (2019)