REFORMULATING UNCALIBRATED PHOTOMETRIC STEREO AS A JOINT DEPTH–NORMAL ESTIMATION TASK

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncalibrated Photometric Stereo (UPS) is traditionally formulated as a surface normal estimation problem, with post-integration to reconstruct depth. This **decoupled** strategy ignores the intrinsic coupling between local shape (normal) and global geometry (depth), leading to inconsistent reconstructions and error accumulation. To address this, we reformulate UPS as a **joint** depth–normal estimation problem that enforces mutual geometric constraints. Our framework leverages CLIP as a stable semantic cues to resolve illumination ambiguity by providing pseudo-lighting cues. A plane-sweeping-like module first recovers a coarse depth map, which is then jointly refined with camera intrinsics. The refined depth and intrinsics are used to derive geometry-consistent surface normals, enabling a consistency-driven supervision that supports semi-supervised learning with limited synthetic data and bridges photometric and geometric reconstruction. Extensive experiments demonstrate that our method outperforms state-of-the-art UPS methods and is competitive with calibrated and depth estimation methods. Ablation studies confirm that CLIP provides robust semantic cues that effectively stabilize the reconstruction process.

1 INTRODUCTION

Photometric stereo (PS) aims to recover surface normals of an object from images captured under varying illumination at a fixed viewpoint. (Woodham, 1980; Ju et al., 2024) The normals are fundamental for downstream tasks such as relighting (Kim et al., 2024), 3D reconstruction (Bansal et al., 2016), and geometric understanding (Zhang et al., 2022). Existing methods fall into two categories: *calibrated* and *uncalibrated*. Calibrated PS (Ikehata et al., 2012; Enomoto et al., 2020) assumes known lighting directions from explicit calibration, while uncalibrated PS (UPS) (Xu et al., 2025; Kaya et al., 2021) estimates both surface normals and illumination directly from observations. Owing to its flexibility and the removal of calibration constraints, UPS has attracted increasing attention in recent years. Despite this progress, existing UPS methods face fundamental limitations in geometric reconstruction. First, most methods **estimate surface normals without depth** and rely on post-integration (Kazhdan et al., 2006), which often fails to recover fine structures. (The reason can be found in Appendix. A.1) This decoupled design breaks the complementarity between local shape (normal) and global geometry (depth), causing error accumulation and surface discontinuities (Fig. 1). Second, **one-dimensional normal supervision** alone underutilizes geometric cues, and the absence of explicit lighting information further limits representation capacity and generalization. To overcome these issues, we reformulate UPS as a *joint depth–normal estimation* problem, where mutually constrained optimization enforces geometric consistency and enables holistic reconstruction under uncalibrated lighting. Depth, as an essential geometric attribute, should not only serve as an *output* but also as a *supervisory signal* for normal estimation. Leveraging their analytical relation—normals can be derived from depth given camera intrinsics K—we enforce depth–normal consistency to achieve mutual supervision, enabling semi-supervised depth learning during training (since most photometric stereo datasets lack ground-truth depth), which is further refined with limited synthetic ground truth. This geometry-aware formulation bridges the gap in uncalibrated settings and enables coherent recovery of both depth and surface orientation. Specifically, our framework consists of two stages: a *reconstruction stage* and a *geometry-aware learning stage*. In the reconstruction stage, we employ CLIP (Radford et al., 2021) to extract **stable semantic cues** that provide pseudo-calibrated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

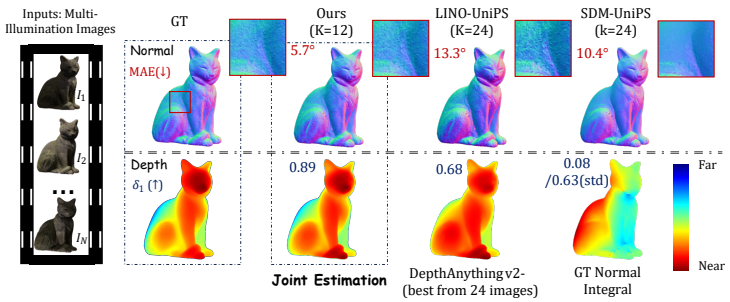


Figure 1: Given multi-illumination images, our method jointly estimates excellent surface normals and depth. The monocular depth estimation baselines are reported using their best result across all input images, which approximates an upper bound for such methods in this setting. Proving that our method can achieve more accurate depth estimation by leveraging complementary cues from multi-illumination inputs, which is important for applications that need precise depth of objects.

lighting information. Contrast-enhanced images are matched with a set of directional text prompts, enabling the network to infer relative illumination directions. These pseudo-light embeddings enrich the representation of uncalibrated photometric stereo. The multi-illumination images are concatenated with the embeddings and fed into an end-to-end network for surface normal estimation. Inspired by ray tracing, we further design a plane-sweeping-like (Nievergelt & Preparata, 1982; Collins, 1996; Gallup et al., 2007) module to aggregate shading information across all illuminations. It constructs a depth cost volume by evaluating each depth hypothesis via the photometric consistency between predicted and observed shading under the inferred lighting geometry, thereby mapping illumination variations to geometric depth cues. In the geometry-aware learning stage, a secondary network refines the coarse depth and photometric features to jointly estimate fine depth and camera intrinsics K. We adopt a simplified camera perspective model by fixing the principal point and only estimating the focal length to simplify the optimization. (The performance under non-ideal camera settings is provided in the Appendix. A.5) The refined depth and intrinsics are used to derive reference normals, which are aligned with the photometric normals via a depth–normal consistency loss. Together with supervised normal loss, this enables depth learning without ground-truth depth while providing complementary guidance for normal reconstruction. Finally, the model is fine-tuned on a small Blender-rendered dataset with GT normal and depth maps for improved accuracy. Consequently, our framework achieves high-quality joint estimation of surface normals and depth under uncalibrated photometric stereo. Our contribution includes:

- a geometry-aware supervision paradigm for photometric stereo. We introduce a depth–normal consistency loss that provides an additional supervision pathway beyond direct normal regression, alleviating the limited supervision diversity in existing methods and improving normal reconstruction quality.
- a plane-sweeping-like module for depth estimation. It constructs a physically grounded depth cost volume by evaluating each depth hypothesis through photometric consistency. By aggregating shading cues from multi-illumination images, it provides a reliable geometric prior for accurate and consistent depth reconstruction.
- a unified UPS pipeline and synthetic dataset for joint depth–normal estimation. Compared to monocular depth estimation methods, our approach utilizes multi-illumination inputs to achieve more accurate depth estimation, which is crucial for applications that require precise depth of objects. We further construct a Blender-rendered dataset with detailed objects and ground-truth depth and normals for fine-tuning and evaluation.

2 Related Works

2.1 Calibrated Photometric Stereo (CPS)

Classical photometric stereo (PS) (Woodham, 1980) assumes Lambertian reflectance and known lighting directions, enabling shape recovery via linear least-squares optimization. Subsequent

works extended this framework to handle non-Lambertian materials through robust statistical (Wu et al., 2010; Hui & Sankaranarayanan, 2016; Hertzmann & Seitz, 2005) or parametric reflectance models (Ikehata & Aizawa, 2014; Shi et al., 2013). With the advent of deep learning, data-driven CPS (Chen et al., 2018; Honzátko et al., 2021; Yao et al., 2020) methods have achieved state-of-the-art normal estimation by leveraging global feature aggregation across multiple illuminations, including CNN- (Santo et al., 2017; Zheng et al., 2019) and Transformer-based (Ju et al., 2022; Liu et al., 2020; Ikehata, 2022a; Vaswani et al., 2017) designs. Despite these advances, they all rely on pre-calibrated lighting and thus cannot generalize to uncalibrated scenarios.

2.2 Uncalibrated Photometric Stereo (UPS)

Uncalibrated photometric stereo (UPS) removes the need for known lighting, offering greater flexibility for real-world use. Classical optimization-based UPS methods address the geometry–lighting ambiguity (e.g., the generalized bas-relief ambiguity) (Belhumeur et al., 1999) through additional priors such as inter-reflections (Chandraker et al., 2005), specular spikes (Drbohlav & Šára, 2002), or parameteric specular reflection (Georghiades, 2003). However, these methods rely on strong assumptions or manual annotations, which limit their practicality. Recent learning-based approaches have advanced UPS by jointly estimating lighting and surface normals through neural networks. Two-stage frameworks (Chen et al., 2019; 2020) first infer lighting directions and then estimate normals, while end-to-end networks learn both simultaneously. Some studies (Kaya et al., 2021; Li & Li, 2022b) further integrate differentiable inverse rendering to refine lighting and geometry, yet most still require explicit lighting estimation and produce only normal maps rather than complete geometry.

2.3 Depth Estimation in Photometric Stereo

Research on direct depth estimation within the photometric stereo framework remains limited. Most existing studies focus on normal estimation, obtaining depth indirectly through integration. A few works attempt to predict depth under specific assumptions. For example, variational uncalibrated photometric stereo (Haefner et al., 2019) jointly optimizes depth and lighting using a Lambertian reflectance model and low-order spherical harmonics. These methods perform well on ideal diffuse surfaces but are inherently model-dependent and lack generality for complex reflectance or unconstrained lighting. Similarly, optimization-based inverse rendering (Peng et al., 2017) incorporates photometric cues to refine depth but often depends on auxiliary inputs or strong priors absent in standard photometric stereo. In summary, a general and model-agnostic framework for jointly recovering depth and normals from multi-illumination images—especially under uncalibrated or weakly supervised conditions—remains missing.

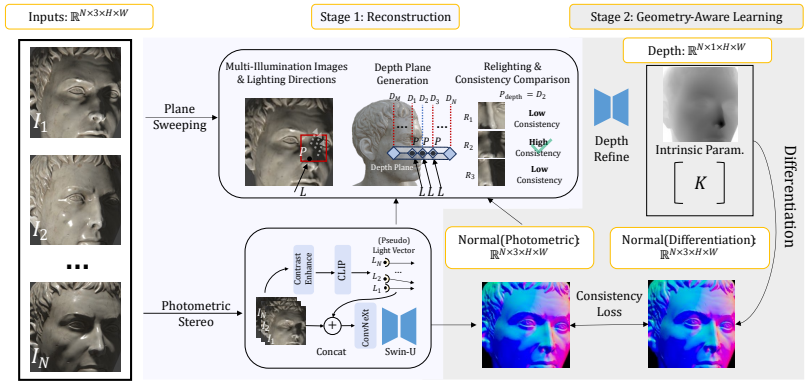
3 Method

3.1 Overview

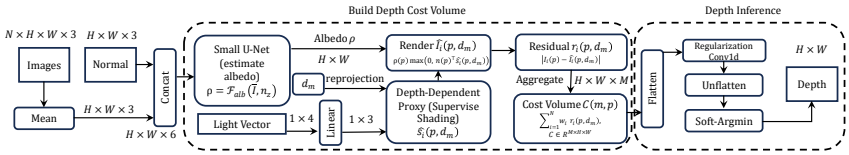
We address the *uncalibrated photometric stereo* (UPS) problem by jointly estimating surface normals and depth from images captured under varying illumination. The input is a set of N multi-illumination RGB images $I \in \mathbb{R}^{N\times 3\times H\times W}$, and the outputs are the surface normal map $N \in \mathbb{R}^{3\times H\times W}$ and depth map $D \in \mathbb{R}^{1\times H\times W}$. As shown in Fig. 2, our framework consists of two stages: a *reconstruction stage* and a *geometry-aware learning stage*. The former uses *pseudo-calibrated photometric stereo* for normal estimation and *plane sweeping* for coarse depth recovery, while the latter refines depth via a *Depth Refine Module* that also estimates camera intrinsics K. Finally, *joint supervision through geometric consistency* optimizes both predictions, enabling depth learning without ground-truth supervision and improving normal accuracy.

3.2 Reconstruction Stage

We first estimate per-pixel surface normals from multi-illumination inputs and then recover a coarse depth map via a plane-sweeping strategy. The predicted normal and depth are denoted as $\hat{N} \in \mathbb{R}^{N\times 3\times H\times W}$ and $\hat{D} \in \mathbb{R}^{N\times 1\times H\times W}$, respectively.



(a) Overview of our proposed geometry-aware joint depth–normal estimation framework under uncalibrated photometric stereo. The pipeline consists of two main stages: a *reconstruction stage* and a *geometry-aware learning stage*.



(b) Detailed view of the plane-sweeping module. The core idea of this module is to approximate ray tracing and to determine the most probable depth by comparing photometric consistency.

Figure 2: Pipeline of our proposed method.

3.2.1 Photometric Stereo

Output: photometric normal $N_{\text{photo}} \in \mathbb{R}^{3\times H\times W}$. We first leverage CLIP to extract **stable semantic lighting cues**. For each illumination image I_n, its CLIP feature is matched with a set of directional prompts $\mathcal{T} = \{\text{left}, \text{right}, \text{top}, \text{bottom}\}$ (full descriptions are provided in the Ablation Study section), yielding the semantic distribution $s_n = \text{softmax}\left(\frac{f_{\text{img}}(I_n) f_{\text{text}}(\mathcal{T})^\top}{\tau}\right)$, $s_i = Linear(s_n, 5 \to 3)$, where f_{img} and f_{text} are CLIP encoders and τ is a learnable temperature. Each s_n is broadcast to the image resolution and concatenated with I_n. The network then decodes the concatenated features into N_{photo} using a Swin-UNet backbone (Liu et al., 2021; Cao et al., 2022), which combines hierarchical self-attention with multi-scale feature fusion.

3.2.2 Plane-Sweep-like Reconstruction

Output: coarse depth $D_{\text{coarse}} \in \mathbb{R}^{1\times H\times W}$. We adopt a *plane-sweep-like* formulation to infer depth from single-view multi-illumination shading by constructing a depth-dependent photometric cost volume.

Depth-dependent photometric model. Under an orthographic assumption, a pixel $\mathbf{p} = (x, y)$ with depth hypothesis d_m is lifted to $\mathbf{P}(\mathbf{p}, d_m) = (x, y, d_m)^\top$. Given surface normal $n(\mathbf{p})$ and light direction $s_i \in \mathbb{R}^3$, the depth-dependent incident direction is

$$\hat{s}_i(\mathbf{p}, d_m) = \frac{s_i - \alpha\, \mathbf{P}(\mathbf{p}, d_m)}{\|s_i - \alpha\, \mathbf{P}(\mathbf{p}, d_m)\|_2}, \tag{1}$$

where α is a learnable parameter initialized to 0.1. Since subsequent network refinement can optimize the representation, we use the Lambertian model to simplify the computation. The predicted intensity:

$$\hat{I}_i(\mathbf{p}, d_m) = \rho(\mathbf{p})\ \max\big(0,\ n(\mathbf{p})^\top \hat{s}_i(\mathbf{p}, d_m)\big)\,, \tag{2}$$

where $\rho(\mathbf{p})$ is the albedo and estimated by a small U-Net. The residual $r_i(\mathbf{p}, d_m) = \left|I_i(\mathbf{p}) - \hat{I}_i(\mathbf{p}, d_m)\right|$. Each depth hypothesis thus induces a distinct shading prediction, making the residual explicitly depth-dependent.

Depth cost volume. The depth-dependent residuals are aggregated to form the core of our plane-sweep-like formulation, in which geometric hypotheses are directly evaluated by shading consistency.

The depth range is discretized into M planes $\{d_m\}_{m=1}^{M}$ and residuals are aggregated over N illuminations, with learnable weights w_i to adaptively balance contributions from different lighting:

$$C(m, \mathbf{p}) = \sum_{i=1}^{N} w_i \, r_i(\mathbf{p}, d_m), \tag{3}$$

yielding a depth-dependent cost volume $C \in \mathbb{R}^{M \times H \times W}$.

Depth inference. The cost volume is weakly regularized along the depth axis,

$$\tilde{C}(m, \mathbf{p}) = C(m, \mathbf{p}) + \big(\mathcal{F}_{\text{reg}} * C(\cdot, \mathbf{p})\big)(m), \tag{4}$$

where $\mathcal{F}_{\text{reg}}$ is a 1D convolution kernel, and converted into a probability volume via softmin:

$$P_m(\mathbf{p}) = \frac{\exp(-\tilde{C}(m, \mathbf{p}))}{\sum_{k=1}^{M} \exp(-\tilde{C}(k, \mathbf{p}))}. \tag{5}$$

The final depth is the expectation

$$\hat{D}_{\text{coarse}}(\mathbf{p}) = \sum_{m=1}^{M} P_m(\mathbf{p}) \, d_m. \tag{6}$$

3.3 Geometry-Aware Learning Stage

In this stage, we refine the coarse depth and jointly optimize depth and surface normals under a geometry-aware framework, even when only ground-truth normals are available.

3.3.1 Depth Refine Module

Output: Depth $D \in \mathbb{R}^{1 \times H \times W}$. The plane-sweeping stage produces a coarse depth map whose precision is limited by the discretization of M planes: $\Delta D \approx \frac{D_{\max} - D_{\min}}{2M}$. To overcome this limitation, we introduce a lightweight CNN-UNet–based *Depth Refine Module*. It takes the coarse depth $\hat{D}_{\text{coarse}}$ and its confidence map P_m as input, and aligns the encoded features with those from the photometric stereo stage. The concatenation enables cross-modal interaction between photometric and geometric cues for refined depth estimation.

3.3.2 Geometric Consistency Supervision

Since most photometric stereo datasets lack ground-truth depth, we supervise depth implicitly via geometric consistency between depth and normals. Given the refined depth D and intrinsic matrix K, we compute the differential normal N_{diff} from local surface gradients:

$$N_{\text{diff}} = \frac{K^{-1} \, [-\partial_x D, \, -\partial_y D, \, 1]^\top}{\| K^{-1} \, [-\partial_x D, \, -\partial_y D, \, 1]^\top \|}, \tag{7}$$

where $\partial_x D$ and $\partial_y D$ are spatial depth derivatives, yielding a differentiable geometry-induced normal field. (We use camera perspective model to simplify the optimization, and the performance under non-ideal camera settings is provided in the Appendix. A.5)

To enforce mutual consistency among depth, photometric normals, and ground-truth normals, we define the joint loss:

$$\mathcal{L}_{\text{joint}} = \sum_{(i,j) \in (\text{photo,diff,gt})} \text{MSE}(N_i, N_j). \tag{8}$$

This formulation enables semi-supervised depth learning while jointly refining photometric normals through cross-consistency regularization.

Table 1: Results on the DiLiGenT benchmark. We report the mean angular error (MAE) in degrees for surface normal estimation. The best results are highlighted in **bold**.

Method	Task	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Ave
LL22a (Li & Li, 2022a)	Calibrated	2.4	3.6	8.0	4.9	4.7	6.7	14.9	6.0	5.0	8.8	6.5
JS22 (Ju et al., 2022)		2.9	5.5	7.1	4.7	6.0	7.5	12.3	6.0	6.4	9.9	6.8
CH19 (Ikehata, 2018)	Uncalibrated	2.8	6.9	9.0	8.1	8.5	11.9	17.4	8.1	7.5	14.9	9.5
IK22 (Ikehata, 2022b)		4.9	9.1	19.4	13.0	11.6	24.2	25.2	10.8	9.9	18.8	14.7
GR23 (Ju et al., 2023)		2.2	5.6	6.7	4.3	6.1	6.8	12.0	5.5	6.4	9.7	6.6
IK23 (Ikehata, 2023)		1.5	3.6	7.5	5.4	4.5	8.5	10.2	4.7	4.1	8.2	5.8
LC25 (Li et al., 2025)		1.8	2.6	**6.2**	3.4	4.3	5.2	8.6	**4.1**	4.6	6.7	4.7
KK21 (Kaya et al., 2021)	Self-Calibrated	3.8	6.0	13.1	7.9	10.9	11.9	25.5	8.8	10.2	18.2	11.6
LL22b (Li & Li, 2022b)		**1.2**	3.8	9.3	4.7	5.5	7.1	14.6	6.7	6.5	10.5	7.1
Ours	Uncalibrated	1.6	**3.2**	6.3	**3.1**	**3.7**	**4.5**	**7.4**	4.3	**3.8**	**6.4**	**4.4**
Ours (K=16)		1.6	3.2	6.3	3.1	3.8	4.6	7.4	4.6	3.9	6.4	4.5
Ours (K=8)		1.6	4.1	7.8	3.5	4.2	5.1	8.3	4.7	4.0	7.1	5.0

4 RESULTS

4.1 TRAINING AND EVALUATION SETTING

4.1.1 TRAINING DETAILS

Following many previous works, we train our model on the PS-Blooby and PS-Sculpture datasets(Chen et al., 2018). In addition, we render about 2000 synthetic samples using Blender, which cover only about 2.3% of the category diversity of PS-Blooby, as an additional dataset for depth fine-tuning. The Hyper parameters of network and optimization can be found in the appendix. A.3.1

4.1.2 EVALUATION SETTING

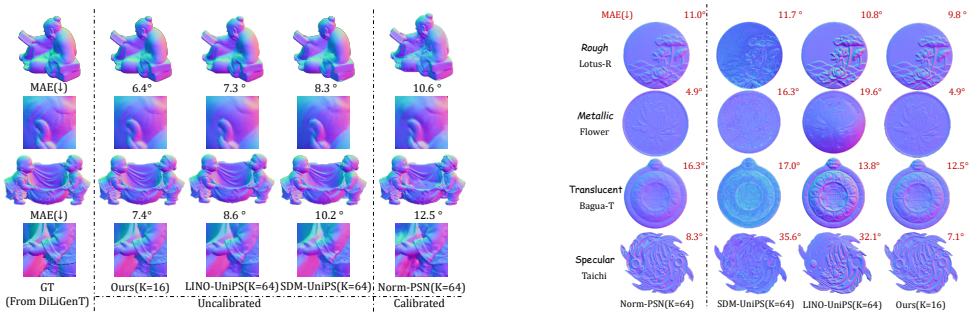
All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU. We compare our method on the DiLiGenT (Shi et al., 2016) and DiLiGenT-Pi (Wang et al., 2023) benchmarks. For methods that accept masks, we provide mask inputs during evaluation. All baselines are tested with as many input images as possible for a fair comparison. For reconstructed depth maps, we evaluate our results both on our synthetic dataset and on the DTU dataset.

4.2 EVALUATIONS ON DILIGENT BENCHMARK

We evaluate our method on the DiLiGenT benchmark, which contains 10 objects under 96 HDR lighting conditions at 612×512 resolution, with ground-truth normals, masks, and calibrated light directions. Unless specified, all methods use the full 96 images. Masked methods apply the corresponding masks, and all predicted normals are multiplied by the mask before evaluation; since MAE is computed within mask regions, this does not affect comparisons. Table 1 summarizes quantitative results, grouping methods into *Calibrated*, *Uncalibrated*, and *Self-Calibrated*. The "Universal" task generalizes the uncalibrated setting, while self-supervised lighting methods are *Self-Calibrated*. As shown, our method consistently outperforms prior approaches. For challenging objects (*Reading*, *Harvest*) with complex geometry and reflectance, it achieves clear gains. Even with only 16 input images, it maintains SOTA-level performance, showing that CLIP-based pseudo-illumination cues and depth supervision effectively compensate for limited input diversity. Qualitative results in Fig. 3a demonstrate that 16-input reconstructions preserve fine details and outperform methods trained with 64 inputs, highlighting the robustness of both modules.

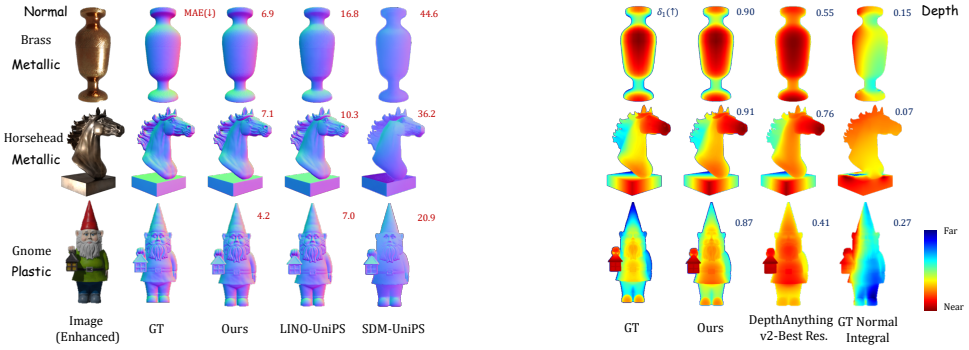
4.3 EVALUATIONS ON DILIGENT-PI BENCHMARK

DiLiGenT-Pi is a large-scale real-world photometric stereo benchmark comprising mostly planar and circular objects, categorized into four material groups: metallic, rough, translucent, and specular.



(a) Comparison on DiLiGenT. (b) Comparison on DiLiGenT-Pi.

Figure 3: Subjective comparison on the DiLiGenT and DiLiGenT-Pi benchmark. Our method recovers more surface details and achieves higher fidelity.



(a) Normal comparison. (b) Depth comparison.

Figure 4: Subjective comparison on the N&D Syn benchmark. Our method jointly recovers high-quality surface normals and depth maps.

Non-Lambertian materials—especially metallic ones—exhibit strong specular highlights that pose major challenges for photometric stereo, often leading to inaccurate normal estimation even on flat regions. We evaluate our method on DiLiGenT-Pi under settings consistent with those of DiLiGenT. The results are summarized in Table 3b.

Appendix. A.4 presents the quantitative results on DiLiGenT-Pi. Our method achieves leading or comparable performance to state-of-the-art approaches across all material categories and even surpasses several calibrated methods in quantitative accuracy. The qualitative comparisons in Fig. 3b further highlight the superior detail recovery of our approach. Across representative examples from all four material groups, our results exhibit accurate surface normals and well-preserved fine geometric details, demonstrating strong robustness and generalization.

4.4 Evaluations of Normals and Depths

4.4.1 Joint Estimation on Synthetic Benchmark

To assess the joint reconstruction capability of our method, we built a synthetic benchmark in Blender, named the N&D Syn, containing five objects with diverse surface properties(smooth, rough, detailed, and highly reflective metallic materials). Each object was rendered under 24 point-light illumination conditions at a resolution of 512×512. For comparison, we included representative monocular depth estimation methods, which take single-image inputs. To ensure fairness, we report their best performance among the 24 rendered images for each object. Qualitative and quantitative results are presented in Fig. 4 and Table 2. In Table 2, "GT Normal Integral" denotes depth reconstructed by integrating ground-truth normals via Poisson integration, while "GT Normal Integral (std.)" represents its normalized version to remove global scale and offset ambiguities. Our method accurately reconstructs both normals and depth, outperforming all monocular methods—even their best cases—demonstrating the advantage of multi-illumination cues for fine geometric recovery.

Table 2: Quantitative comparison on the N&D Syn and the DTU datasets. The other methods take 24 images as inputs. The DTU datasets only have depth GT, so we only report depth results.

	Brass	Cat	Duck	Horsehead	Gnome	Ave	Scan6	Scan18
Depth ($\delta_1 \uparrow$)								
DepthAnythingv2	0.31	0.49	0.32	0.58	0.10	0.36	0.65	0.47
DepthAnythingv2† (Yang et al., 2024)	0.55	0.65	0.48	0.76	0.41	0.57	0.72	0.49
Metric3d (Hu et al., 2024)	0.29	0.40	0.09	0.41	0.13	0.26	0.65	0.57
Metric3d†	0.39	0.55	0.14	0.51	0.19	0.36	0.75	0.64
VideoDepthAnything (Chen et al., 2025)	0.30	0.48	0.25	0.59	0.10	0.34	0.52	0.68
GT Normal Integral	0.15	0.082	0.096	0.071	0.27	0.13	/	/
~(std.)	0.64	0.63	0.68	0.61	0.59	0.63	/	/
Ours(K=12)	**0.90**	**0.88**	**0.92**	**0.91**	**0.87**	**0.90**	**0.84**	**0.85**
Normal (Mean Angular Error)↓								
SDM-UniPS	16.8	10.4	8.3	36.2	20.9	18.5	/	/
LINO-UniPS	44.6	13.3	5.2	10.3	7.0	16.1	/	/
Ours	**6.9**	**5.7**	**4.1**	**7.1**	**4.2**	**5.6**	/	/

Table 3: Ablation study on depth and normal estimation.

Normal Estimation (MAE ↓)			**Depth Estimation ($\delta_1 \uparrow$)**		
Method	DiLiGenT	DiLiGenT-Pi	Method	N&D Syn	DTU
Full model	**4.4**	**8.3**	Full model	**0.90**	**0.85**
w/o Depth sup.	5.3	15.7	w/o Normal sup.	0.67	0.54
w/o CLIP	5.6	13.5	w/o Plane sweep.	0.79	0.75

Moreover, directly integrating ground-truth normals produces inferior and unstable depth (Fig. 4), underscoring the necessity of our joint reconstruction design. It is worth noting that monocular depth baselines are reported with their best result over all input images, approximating their upper-bound performance. Our method still achieves higher accuracy by exploiting multi-illumination cues, which have more pracical value in the scene that need precise depth estimation of objects.

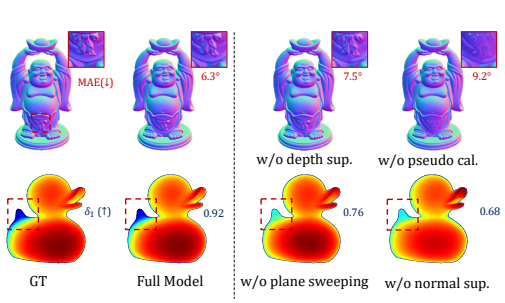
4.4.2 Depth Estimation on Real-World Dataset

To evaluate our method on real-world data, we conducted experiments on a subset of the DTU dataset, as shown in Table 2. DTU scenes are captured under complex, non-point lighting, deviating from the ideal photometric stereo assumption. Despite being fine-tuned on a small point-light dataset, our model achieves significant gains over existing monocular depth estimation methods. This improvement does not result from dataset coupling but from the model's geometry-consistent design and the robustness of the CLIP-guided pseudo-calibration, which generalize effectively beyond the training illumination setup. Our results exhibit finer geometric details and stable depth under non-ideal lighting, validating both the efficacy of the fine-tuning strategy and the intrinsic generalization ability of the network.

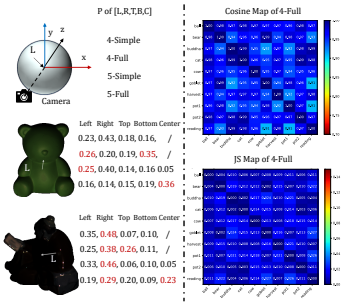
4.5 Ablation study

4.5.1 Components Ablation

We evaluate the contribution of each major component via four targeted ablations focusing separately on depth and normal estimation. For depth, we replace the plane-sweep module with normal integration ("w/o plane-sweep") and we remove the self-supervision arising from normal-to-depth consistency, keeping only supervised fine-tuning ("w/o normal sup."). For normals, we replace CLIP-based pseudo-lighting with $0^{1\times 5}$ ("w/o CLIP") and we remove the depth–normal consistency loss ("w/o depth sup."). Table 3 and Fig. 5a report averaged quantitative results and representative qualitative comparisons on several sets. Results show that (i) both plane-sweeping and normal-guided supervision materially improve depth accuracy, and (ii) pseudo-calibration from CLIP boosts normal quality in high-detail regions while depth supervision further benefits normals on near-planar benchmarks (DiLiGenT-Pi).



(a) Ablation study on depth and normal estimation. Each component contributes to improved reconstruction quality.



(b) Ablation study on different CLIP prompt designs. 4-Full provides the most consistent semantics.

Figure 5: Ablation study on depth and normal estimation, and analysis of CLIP prompt designs.

Table 4: Comparison of different CLIP prompt designs. Directional accuracy is not strictly correlated with reconstruction quality, while cross-object statistical consistency shows a strong correlation.

	Cosine(↑)	JS(↓)	Accuracy (%)	MAE(↓)	δ_1(↑)
4-Simple	0.92	0.028	74.0	4.49	0.89
4-Full	**0.96**	**0.011**	**80.2**	**4.35**	**0.90**
5-Simple	0.88	0.039	55.2	4.57	0.89
5-Full	0.93	0.019	64.3	4.41	0.90

4.5.2 CLIP Prompts

We analyze how different CLIP-prompts affect the network by examining both their statistical properties and their impact on reconstruction. Two factors are considered: (1) directional categories and (2) prompt complexity. Specifically, **4-Simple** uses concise prompts (e.g., *bright from left/right/top/bottom*), **5-Simple** additionally includes *bright from center*, and **4-Full** extends each direction with detailed descriptions (e.g., *light source from the left relative to the camera, left side of the object significantly brighter*). For fair comparison, the input channel of the first bottleneck layer is fixed to 5; for 4-direction prompts, the fifth channel is zero-padded. Table 4 and Fig. 5b report the results. Accuracy measures whether the two highest CLIP probabilities match the ground-truth directions, while Cosine and JS divergence quantify cross-object consistency of CLIP statistics on DiLiGenT. MAE and δ_1 evaluate normal and depth accuracy on DiLiGenT and N&D Syn, respectively. **The semantic correctness of CLIP directional cues is not strongly correlated with reconstruction quality.** Instead, **cross-object stability of CLIP statistics (Cosine ↑, JS ↓) shows a much stronger correlation with reconstruction performance**, consistent with our design of using CLIP as a *stable semantic prior* rather than an accurate lighting estimator. Moreover, detailed prompts with clearly separated semantics further improve statistical stability. Overall, **4-Full** provides the most consistent semantics and achieves the best reconstruction quality, and is adopted in our model.

5 Conclusion

We presented a unified framework for joint depth and normal estimation in UPS. By enforcing depth–normal consistency and introducing CLIP-based pseudo calibration with plane sweeping, our method achieves coherent and accurate geometry reconstruction under uncalibrated lighting. Extensive experiments confirm its superior performance and robustness across multiple benchmarks. Although accuracy remains high under universal lighting (e.g., DTU), performance slightly drops compared with ideal point-light setups due to limited training diversity. Future work will explore broader illumination variations to further improve generalization and real-world applicability.

AI USE STATEMENT

Generative AI tools were used for drafting, editing, and code-bootstrap assistance during manuscript and package preparation. All experiments, statistics, and conclusions were designed, executed, and verified by human authors; AI-assisted text was reviewed and corrected by the authors, who take responsibility for the final content.

ETHICS STATEMENT

The study uses only public photometric-stereo datasets (PS-Blobby, PS-Sculpture) and the public DTU benchmark, together with a Blender-rendered synthetic dataset created by the authors; no human subjects and no personal or demographic data are involved. The work concerns object-level geometry reconstruction from images. As with other geometry-estimation methods, the reconstructed depth and normals could in principle be repurposed for unconsented reconstruction of real objects or scenes; the method is developed and evaluated only on object-level academic benchmarks, and this scope is stated in the paper.

REPRODUCIBILITY STATEMENT

Architecture details, the construction of the Blender-rendered synthetic dataset, and the training and evaluation settings are fully specified in the main text and the technical appendices. Source code and the synthetic dataset will be released upon publication.

REFERENCES

Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5965–5974, 2016.

Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International journal of computer vision*, 35(1):33–44, 1999.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Manmohan Krishna Chandraker, Fredrik Kahl, and David J Kriegman. Reflections on the generalized bas-relief ambiguity. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 788–795. IEEE, 2005.

Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–18, 2018.

Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8739–8747, 2019.

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

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025.

Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition*, pp. 358–363. Ieee, 1996.

Ondřej Drbohlav and Radim Šára. Specularities reduce ambiguity of uncalibrated photometric stereo. In *European Conference on Computer Vision*, pp. 46–60. Springer, 2002.

Kenji Enomoto, Michael Waechter, Kiriakos N Kutulakos, and Yasuyuki Matsushita. 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.

David Gallup, Jan-Michael Frahm, Philippos Mordohai, Qingxiong Yang, and Marc Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2007.

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.

Bjoern Haefner, Zhenzhang Ye, Maolin Gao, Tao Wu, Yvain Quéau, and Daniel Cremers. Variational uncalibrated photometric stereo under general lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8539–8548, 2019.

Aaron Hertzmann and Steven M Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (8):1254–1264, 2005.

David Honzátko, Engin Türetken, Pascal Fua, and L Andrea Dunbar. Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo. In *2021 International Conference on 3D Vision (3DV)*, pp. 394–402. IEEE, 2021.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhuo Hui and Aswin C Sankaranarayanan. Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10): 2060–2073, 2016.

Satoshi Ikehata. 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.

Satoshi Ikehata. Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism. *arXiv preprint arXiv:2211.11386*, 2022a.

Satoshi Ikehata. Universal photometric stereo network using global lighting contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12591–12600, 2022b.

Satoshi Ikehata. 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.

Satoshi Ikehata and Kiyoharu Aizawa. Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2179–2186, 2014.

Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa. Robust photometric stereo using sparse regression. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 318–325. IEEE, 2012.

Yakun Ju, Boxin Shi, Muwei Jian, Lin Qi, Junyu Dong, and Kin-Man Lam. Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention. *International Journal of Computer Vision*, 130(12):3014–3034, 2022.

Yakun Ju, Boxin Shi, Yang Chen, Huiyu Zhou, Junyu Dong, and Kin-Man Lam. Gr-psn: Learning to estimate surface normal and reconstruct photometric stereo images. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

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

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

Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25096–25106, 2024.

Hong Li, Houyuan Chen, Chongjie Ye, Zhaoxi Chen, Bohan Li, Shaocong Xu, Xianda Guo, Xuhui Liu, Yikai Wang, Baochang Zhang, et al. Light of normals: Unified feature representation for universal photometric stereo. *arXiv preprint arXiv:2506.18882*, 2025.

Junxuan Li and Hongdong Li. Neural reflectance for shape recovery with shadow handling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16221–16230, 2022a.

Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *European Conference on Computer Vision*, pp. 166–183. Springer, 2022b.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Jürg Nievergelt and Franco P Preparata. Plane-sweep algorithms for intersecting geometric figures. *Communications of the ACM*, 25(10):739–747, 1982.

Songyou Peng, Bjoern Haefner, Yvain Quéau, and Daniel Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 2961–2968, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo network. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 501–509, 2017.

Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE transactions on pattern analysis and machine intelligence*, 36(6): 1078–1091, 2013.

Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Feishi Wang, Jieji Ren, Heng Guo, Mingjun Ren, and Boxin Shi. Diligent-pi: A photometric stereo benchmark dataset with controlled shape and mate-rial variation. In *ICCV*, 2023.

Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.

Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian conference on computer vision*, pp. 703–717. Springer, 2010.

Minzhe Xu, Xin Ding, You Yang, Yinqiang Zheng, and Qiong Liu. A Serial Perspective on Photometric Stereo of Filtering and Serializing Spatial Information. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–15, 2025. doi: 10.1109/TVCG.2025.3546657.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, and Boxin Shi. Gps-net: Graph-based photometric stereo network. *Advances in Neural Information Processing Systems*, 33:10306–10316, 2020.

Jingbo Zhang, Ziyu Wan, and Jing Liao. Adaptive joint optimization for 3d reconstruction with differentiable rendering. *IEEE Transactions on Visualization and Computer Graphics*, 29(6): 3039–3051, 2022.

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

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 WHY NORMAL INTEGRATION FAILS EVEN WITH GROUND-TRUTH NORMALS

A.1.1 RELATION BETWEEN NORMALS AND DEPTH

Let $\mathbf{n} = (n_x, n_y, n_z)^\top$. Under orthographic projection,

$$\mathbf{n} \propto (-\partial_x z, -\partial_y z, 1)^\top, \tag{9}$$

so the implied gradients are

$$p = -n_x/n_z, \qquad q = -n_y/n_z. \tag{10}$$

A valid depth z exists only when the *integrability* condition holds:

$$\partial_y p = \partial_x q. \tag{11}$$

A.1.2 WHY INTEGRATION FAILS

Depth is usually recovered by the Poisson least-squares formulation:

$$\min_z \iint_\Omega \left((\partial_x z - p)^2 + (\partial_y z - q)^2\right) dx\, dy, \tag{12}$$

which yields the Poisson equation:

$$\nabla^2 z = \partial_x p + \partial_y q. \tag{13}$$

Even with ground-truth normals, the reconstructed depth exhibits bias and global distortion due to the following intrinsic issues.

1. Perspective–orthographic mismatch. Ground-truth normals are defined under perspective projection, where the normal–gradient relation is nonlinear:

$$\partial_x z \neq -\frac{n_x}{n_z}, \qquad \partial_y z \neq -\frac{n_y}{n_z}. \tag{14}$$

Hence the gradients computed from $\mathbf{n}_{\mathrm{GT}}$ violate (11), making perfect integration impossible.

2. Discrete non-integrability. Even if the surface is smooth, the sampled normal field on a discrete grid almost never satisfies $\partial_y p = \partial_x q$ exactly. Small finite-difference inconsistencies accumulate globally, causing the Poisson solver to produce a "best-fit" surface rather than the true depth.

3. Harmonic ambiguity. Solutions of (13) are defined only up to an additive harmonic function:

$$z' = z + h, \qquad \nabla^2 h = 0. \tag{15}$$

This null space includes global shift, tilt, and other low-frequency bending, so the mapping $\mathbf{n} \mapsto z$ is not unique. **Summary.** Even ground-truth normals fail to yield accurate depth because the normal field, as shown in Fig. 6,

- does not satisfy orthographic integrability due to perspective geometry,
- cannot be perfectly integrable after discretization,
- determines depth only up to harmonic ambiguity.

A.1.3 WHY INTEGRATED DEPTH REQUIRES GROUND-TRUTH NORMALIZATION

The Poisson reconstruction in Eq. (13) determines depth only up to an additive *harmonic function*. Specifically, if z minimizes Eq. (12), then for any h satisfying $\nabla^2 h = 0$, the function $z' = z + h$ produces the same gradient field (p, q) and is therefore an equally valid solution. This null space includes global shift, planar tilt, and other low-frequency distortions.

Thus the recovered depth is defined only up to the equivalence class $z \sim z + h$,$h \in \ker(\nabla^2)$ which implies that its scale and offset cannot be fixed by normal integration alone. In practice, integrated depth must be aligned to ground-truth depth (e.g., by matching mean/variance) before quantitative evaluation.

In contrast, a neural network directly outputs depth in a fixed numerical range (e.g., through bounded activations), eliminating this harmonic ambiguity. Therefore, normal integration inherently contains additional degrees of freedom that learned depth estimation does not.

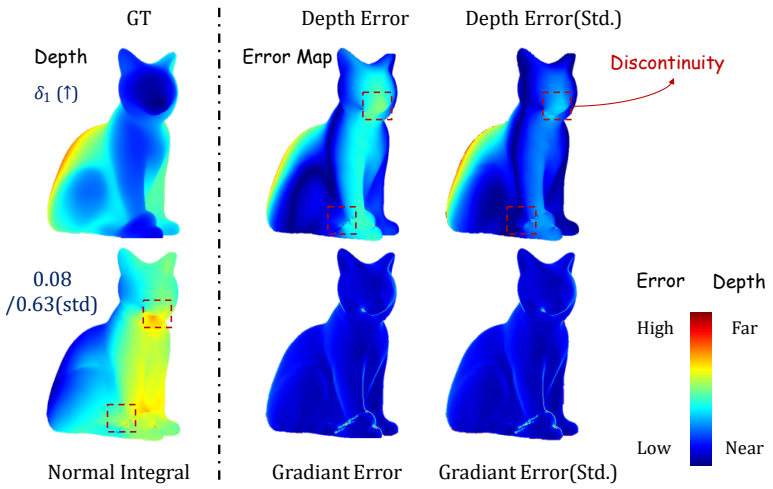


Figure 6: Visualization of depth reconstructed from ground-truth normals. For the gradient error, values are scaled using percentile-based clipping (1%–99%) to enhance contrast and highlight discontinuous regions

A.2 Detailed Analysis of CLIP Statistical Consistency

As discussed in the main paper, we conduct a detailed analysis of CLIP statistics under different prompt designs by measuring the cross-object cosine similarity and Jensen–Shannon (JS) divergence on the DiLiGenT benchmark. These metrics characterize the consistency of CLIP-extracted semantic lighting distributions across different objects under the same illumination direction.

Given two CLIP probability distributions p and q corresponding to the same lighting direction but different objects, cosine similarity and JS divergence are defined as

$$\text{Cosine}(p, q) = \frac{p^\top q}{\|p\|_2 \, \|q\|_2}, \tag{16}$$

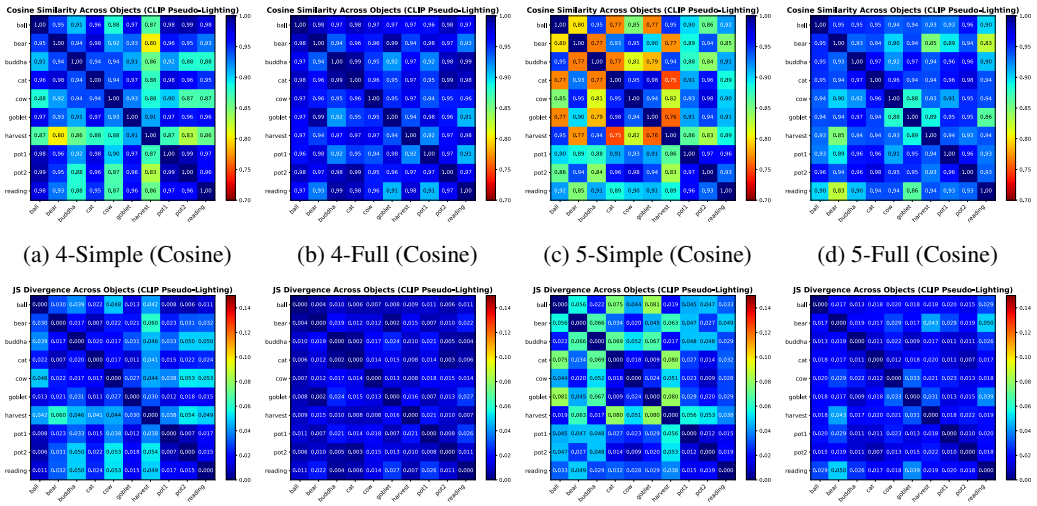
$$\text{JS}(p, q) = \frac{1}{2}\text{KL}(p\|m) + \frac{1}{2}\text{KL}(q\|m), \quad m = \frac{1}{2}(p + q), \tag{17}$$

where $\text{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence.

We compare **4(5)-Simple** prompts, which use concise directional descriptions (*bright from left/right/top/bottom/(center)*), with **4(5)-Full** prompts that provide explicit and detailed semantic descriptions of illumination direction, e.g., *"light source from the (direction) relative to the camera, (direction) side of the object significantly brighter."* The resulting cosine similarity and JS divergence heatmaps are shown in Fig. 7.

Quantitative Observations. Several consistent trends can be observed from the statistical visualizations:

- **Detailed prompts significantly improve cross-object consistency.** Compared to Simple prompts, Full prompts yield higher cosine similarity and lower JS divergence across most objects, indicating more stable semantic representations under varying reflectance.
- **Improvements are most pronounced on objects with complex reflectance.** In DiLiGenT, *Reading* (highly glossy ceramic) and *Harvest* (mixed plastic and metallic materials) exhibit the strongest non-Lambertian effects. For these objects, Full prompts substantially reduce inter-object statistical variance, demonstrating enhanced robustness to reflectance-induced ambiguity.
- **The "center" direction introduces semantic instability.** CLIP tends to assign disproportionately strong or weak probabilities to the "center" prompt, resulting in biased distributions and reduced consistency. Consequently, **5-Simple** consistently underperforms **4-Simple** across all material types.



(a) 4-Simple (Cosine) (b) 4-Full (Cosine) (c) 5-Simple (Cosine) (d) 5-Full (Cosine)

(e) 4-Simple (JS) (f) 4-Full (JS) (g) 5-Simple (JS) (h) 5-Full (JS)

Figure 7: Cross-object CLIP statistical consistency on DiLiGenT under different prompt designs. Top row: cosine similarity; bottom row: JS divergence. Detailed prompts (Full) yield more stable cross-object semantics, while including the "center" direction degrades consistency.

- **Semantic precision outweighs category cardinality.** For language-pretrained models such as CLIP, accurate and explicit prompt descriptions are more critical than increasing the number of semantic categories. Although **5-Full** outperforms **4-Simple**, both settings introduce semantic overlap that degrades output stability compared to **4-Full**.

Based on the above statistical trends, we adopt **4-Full** as the prompt design in all experiments, as it provides the most consistent cross-object semantics and the most stable guidance for downstream reconstruction.

A.3 Model and Dataset Details

A.3.1 Model Details

Our model consists of three main modules—*Photometric Stereo*, *Plane Sweep*, and *Depth Refinement*—as shown in Fig. 2. The figure presents the overall architecture; here we provide the key implementation details of each component.

1. Photometric stereo module. This stage predicts per-pixel normals and extracts a compact photometric latent representation. Each input stack of multi-illumination images is first normalized and then fed into two parallel encoders: a ConvNeXt v2 encoder and the encoder part of a Swin-U network. The ConvNeXt v2 encoder processes the input and produces a feature tensor of shape $(B, 98, 768)$ after average-pooling along the illumination dimension ($N = 98$). The Swin-U encoder produces a feature map with the same number of channels. These two features are concatenated channel-wise and passed into the Swin-U decoder. The decoder output is further projected through a sequence of GhostConv layers to recover the predicted surface normals of shape $(B, 3, H, W)$. We denote the photometric latent feature produced by the ConvNeXt encoder as $\mathbf{F}_{\text{photo}} \in \mathbb{R}^{B\times 98\times 768}$, which is used later by the plane-sweep module. The main hyperparameters are:

- **ConvNeXt v2**: input channels 8, embed dimension 96, stage depths $(3, 3, 9, 3)$, output channels 768.
- **Swin-U**: patch size 4, input channels 8, embed dimension 96, stage depths $(2, 2, 2, 2)$.

Under this configuration, the photometric module outputs: (i) the photometric latent feature $\mathbf{F}_{\text{photo}}$ of size $(B, 98, 768)$, and (ii) the photometric normal map $\mathbf{N}_{\text{photo}}$ of size $(B, 3, H, W)$.

2. Plane-Sweep Module Details **Photometric consistency.** Following the formulation in the main paper, our model learns a photometric consistency measure rather than enforcing a strict Lambertian reflectance model. Each CLIP-based pseudo-lighting descriptor 45-dim) is projected into a learnable 3D illumination vector through a linear layer (Linear($4 \to 3$)), followed by ℓ_2 normalization. Per-light confidence is predicted by a separate Linear($5 \to 1$) layer.

The per-pixel albedo $\rho(\mathbf{p})$ is estimated by a lightweight CNN (SmallAlbedoNet) that takes the mean RGB image (averaged over all illuminations) and the photometric normal's n_z component as input. SmallAlbedoNet consists of two 3×3 Conv layers with hidden dimension 32 and a final 1×1 Conv producing a scalar albedo value. These components form the learned photometric consistency cost used in the plane-sweep evaluation.

Depth hypothesis evaluation. The depth range is uniformly discretized into $M{=}64$ hypotheses: $d_1, \ldots, d_{64} \in [0, 1]$. For each depth candidate, the photometric consistency cost is aggregated to construct a cost volume of shape $C \in \mathbb{R}^{64\times H\times W}$. A learnable depth embedding vector $\mathbf{e} \in \mathbb{R}^{64}$ is added to each slice to improve discrimination between similar depth candidates. This cost volume forms the basis for soft depth selection.

Regularization and soft selection. To improve robustness near occlusions and discontinuities, the cost volume is regularized along the depth axis using a 1D convolution: Conv1D$(1, 1, k = 5)$, initialized with descending weights to encourage near-surface consistency. A learnable temperature parameter $\tau = \mathrm{softplus}(\theta)$ controls the sharpness of the soft-argmin normalization used to convert the regularized cost into a probability distribution over the 64 depth candidates. The final coarse depth is obtained as the expectation over this distribution.
This module outputs a coarse depth map D_{coarse} with a shape of *(B, H, W)*.

3. Depth Refinement Module Details The refinement stage fuses the coarse depth $\hat{D}_{\text{coarse}}$, the photometric normal map $\mathbf{N}_{\text{photo}}$, and the photometric latent representation $\mathbf{F}_{\text{photo}}$.

Feature construction. The concatenation of $\hat{D}_{\text{coarse}}$ and $\mathbf{N}_{\text{photo}}$ is processed by several GhostConv layers to produce a 16-channel feature map. In parallel, $\mathbf{F}_{\text{photo}} \in \mathbb{R}^{B\times98\times768}$ is decoded by a separate Swin-U network (independent from the photometric-stereo Swin-U). The decoder output is projected through a GhostBottleneck to another 16-channel map. The two streams are added to obtain the fused refinement feature.

Depth prediction. A final GhostConv layer reduces the fused feature to a single-channel depth map.

Intrinsic estimation. An MLP–Attention block (8 heads) is applied to the fused feature. The MLP halves the channel dimension at each layer (with hidden dimension equal to the average of input/output sizes). The attention output is mapped to a 2-channel vector representing the focal lengths (f_x, f_y), assuming an approximately centered principal point.

Under this configuration, the depth refinement module outputs: (i) the refined depth map $D_{\text{refine}} \in \mathbb{R}^{B\times1\times H\times W}$, and (ii) the estimated intrinsic matrix $K \in \mathbb{R}^{B\times3\times3}$.

Computation Resources. For an input of 12 illumination images at 224×224 resolution, we profile the model using `fvcore` and `nvidia-smi`. The full model requires **115G FLOPs** and contains **65.1M** parameters, corresponding to approximately **243MB** under mixed-precision training. The runtime GPU memory consumption is around **4.1GB**.

Training Details Following many previous works, we train our model on the PS-Blooby and PS-Sculpture datasets. In addition, we render about 2000 synthetic samples using Blender, which cover only about 2.3% of the category diversity of PS-Blooby, as an additional dataset for depth fine-tuning. The network is optimized using the AdamW optimizer with an initial learning rate of 1×10^{-6}, and a step decay schedule that multiplies the rate by 0.8 every 20 epochs. The batch size is set to 8, and the number of input images per batch is randomly sampled from 2 to 24. This design improves the robustness of the model to varying numbers of input images. As described earlier, the total loss is defined in Eq. (12).

A.3.2 DATASET DETAILS

We generate the Ft dataset and the N&D Syn Benchmark using Blender 4.5. All scenes are rendered on a Windows 11 workstation equipped with an Intel Core i7-13700HX CPU and an NVIDIA GeForce RTX 4060 GPU. GPU-accelerated rendering in Blender is not fully stable on this hardware, so we can not report precise runtimes; however, the entire data generation process finishes within 24 hours. All object models are sourced from publicly available assets on the HDRI websites.

Common Settings. To match the illumination characteristics of photometric stereo, we adopt a single point light source for rendering. The light is placed at a distance of 40 m with a nominal power of 40 000 W (Blender light power does not correspond to real-world wattage). Since all objects are smaller than 1 m, the maximum variation in lighting direction between any two surface points is

$$\theta = 2 \arctan\left(\frac{0.5}{40}\right) \approx 1.43^\circ,$$

which is sufficiently small to treat the source as an ideal point light. Lights are randomly distributed on a hemisphere of radius 40 m. We render all images at a resolution of 512×512. Blender outputs normals in world coordinates; with the camera fixed at (90°, 0°, 0°), camera-space normals are obtained by swapping the G/B channels of the Cycles world normal and negating the new B channel. Depth maps are directly taken from Cycles depth output. Figure 8 compares Cycles and EEVEE on a highly reflective object. Cycles produces physically accurate shading, while EEVEE exhibits non-physical artifacts such as spurious self-reflection. These differences primarily appear in non-Lambertian regions and motivate our choice of using Cycles for the more realistic N&D dataset.

Ft Dataset. To increase data diversity, we use a larger focal length so that each render covers only approximately one-sixth of the object height. The camera is shifted downward six times along the z-axis, and the object is rotated by 15° to repeat the process, yielding 144 multi-illumination sets per object. In total, 14 high-quality objects are rendered, resulting in **2,016 multi-light image sets**. The Ft dataset is rendered with the EEVEE engine using 256 samples, and 24 random light directions are captured per scene.

N&D Benchmark. The N&D dataset captures complete objects (with a focal length about 25% of that used for Ft). Its object models do not overlap with Ft. Rendering is performed with the Cycles engine using 512 samples to achieve higher photorealism.

Fine-tuning Procedure. To avoid overfitting to the Ft dataset, we fine-tune the model for 10 epochs with a batch size of 8 and an initial learning rate of 1×10^{-7} using AdamW. The number of illumination images per iteration is randomly sampled between 2 and 24.

The fine-tuning objective is:

$$\begin{aligned}\mathcal{L}_{\text{ft}} = \mathcal{L}_{\text{N}}(\mathbf{N}_{\text{photo}}, \mathbf{N}_{\text{GT}}) + \mathcal{L}_{\text{D}}(D, D_{\text{GT}}) \\ + \ 0.1\, \mathcal{L}_{\text{cons}}(\mathbf{N}_{\text{photo}}, \mathbf{N}_{\text{int}}),\end{aligned} \tag{18}$$

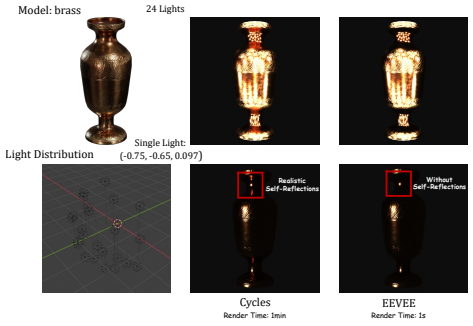


Figure 8: Comparison between Cycles (physically based) and EEVEE (realtime) renderers. Cycles produces physically correct shading, while EEVEE exhibits noticeable non-Lambertian artifacts such as spurious self-reflection (highlighted in red).

Table 5: Rendering settings for the Ft and N&D datasets.

Setting	Ft Dataset	N&D Dataset
Renderer	EEVEE	Cycles
Sampling	256 samples	512 samples
Focal length	Large (∼1/6 object height)	∼25% of Ft
Camera motion	6 vertical shifts	Fixed
Object rotation	15° around z-axis	None
Objects used	14 models	5 models
Sets per object	144	1
Total sets	**2016**	**5**
Shared Settings		
Resolution	512×512	
Lights per scene	24 random directions	
Light sampling	Random point lights on 40 m hemisphere	
Light power	40,000 W (Blender units)	

where the first two terms use weight 1, and the consistency term is scaled by 0.1.

Fine-tuning uses only $1/20$ of the pre-training epochs and 2.3% of the data volume. On a single NVIDIA RTX 4090 GPU with an Intel Xeon Platinum 8331C CPU, the process completes in about 40 minutes.

A.4 Complete DiLiGenT-Pi Results

Table 6: Complete comparison on the DiLiGenT-Pi Benchmark across all object categories. Our method achieves consistently better performance on non-Lambertian objects— particularly in the *Metallic* and *Specular* categories—compared with other UPS-based approaches, demonstrating that the proposed modules remain effective even on challenging non-Lambertian surfaces. Here, **PLANE** denotes the results obtained using an ideal fronto-parallel surface with normal $(0, 0, 1)$. Cell colors follow a blue-to-red colormap: deeper blue indicates better performance, while deeper red corresponds to worse performance.

		Calibrated											Uncalibrated					
	Object	PLANE	LSPS	TH28	TH46	WG10	ST14	PX-Net	CNN-PS	PS-FCN	GPS-Net	NA-PSN	UPS-FCN	SDPS-Net	LW21	SDM-UniPS	Lino-UniPS	Ours
Metallic	BIRD	9	8.1	9.1	8	11.6	10.8	8.8	6.8	7.2	6.8	6.9	15.4	17.6	28.5	26.5	38.3	7.0
	CRAB	5.9	7.1	7.6	8.6	8.5	8.4	7.2	4.5	5.3	4.9	4.4	18.1	20.5	26.8	25.4	39.9	5.0
	FISH	5.2	6.7	7.9	6.5	8.2	8.1	7.4	4.2	4.6	4.6	4.4	15.8	23.6	25.6	24.5	40.0	4.1
	FLOWER	5.8	6.8	7.5	6	7.2	7.9	5.5	4.7	4.6	4.6	4.6	12.7	12.8	23.9	15.2	19.6	5.0
	LIONS	6.6	6.7	7.7	7.4	9.1	8.1	6.8	4.7	4.5	4.6	4.4	11.8	19.6	27.1	9.2	22.0	4.6
	PARA	6.7	6	6.6	5.1	8.4	8	4.9	3.9	4.7	4	4.2	17	19.8	24.6	23.2	18.2	4.4
	QUEEN	6.2	7.5	8.1	8.9	8.6	8.3	14.7	5.4	4.7	4.7	4.9	12.6	16.5	27.2	10.6	36.8	4.6
	RHINO	6.8	7.5	7.8	8.9	9.3	8.9	10.6	4.9	5.3	5.6	5.1	17.7	24.9	26.8	17	9.6	4.9
	SAIL	6.9	8.7	10.2	8.8	8.1	8.2	13.4	5.2	5.1	5.5	5.2	13.5	16.7	27	10.5	20.3	5.9
	SHIP	5.9	7.1	8.9	6.4	9.3	9.1	12.3	4.9	6.1	5.1	4.9	15.6	19	26.4	22	28.0	4.8
Specular	ASTRO	17.6	7.9	8.7	7.9	11.4	9.2	7.1	6	7.2	7.7	7.2	25.9	37.7	36.5	37.8	36.23	6.4
	BEAR	9.8	9.1	8.5	8.9	8.6	9.7	7.9	7.4	7.4	6.9	7.4	27.6	30.7	27	23.8	21.0	6.6
	LUNG	20.3	8.2	8.8	8.1	11.8	9.4	7.6	5.7	9.7	7.5	7.8	36.5	40.2	34.4	46.6	46.8	6.5
	OCEAN	14.5	6.2	6	5.9	8.4	7.6	5.9	4.6	5.8	5.4	5.8	26.7	31.4	28.1	34.6	33.0	4.5
	SUN	11.9	7.9	9.5	11.7	10.5	8.2	6.7	5.8	6.7	8	5.6	22.8	31.5	24	26.2	27.6	5.3
	TAICHI	17.3	9.2	11.2	11	17.9	10.7	8.7	8.3	8	8.3	7.6	29.8	26.9	28.9	36.6	32.7	7.1
	TREE	20.1	9.4	11	9.7	14.9	10.6	11.3	7.8	10.2	9.3	9.7	40.6	34.1	35.8	47.2	41.4	17.9
	TV	19.6	11.7	13.3	13.2	17.1	13.5	12.1	11.3	10.6	10.6	9.6	22.8	41.1	35.4	34.4	32.9	9.1
	WAVE	15.7	6.9	7.7	7.3	9.2	8.4	7.1	5.3	6.8	6.3	6.1	30.8	39.1	25.9	34.9	32.5	5.8
	WHALE	16.3	8.8	8.9	8.5	13.4	10.1	9.5	11.6	12.2	10.4	8.7	37.6	29.8	32	33.8	34.4	6.9
Translucent	BAGUA-T	20.6	16.7	16.5	16.1	17.8	17.4	15.8	16.4	16.8	16.1	16.5	19.9	28.9	27.8	17.1	16.0	12.5
	CLOUD-T	18.7	17	17	16.8	18.1	17.7	16.1	17.2	17.8	17.6	17.3	20.7	27.5	24.2	19.2	16.5	12.9
	LION-T	30.1	21	20.5	19.9	21.6	21.3	18.4	20.3	21.2	23.4	21	25.8	23.6	30.5	16.2	16.3	15.0
	LOTUS-T	15.9	14	14.1	13.8	14.3	14.5	13.7	13.5	13.6	13.7	13.7	19.5	26.5	21.3	13.6	13.8	9.6
	PANDA-T	18.6	16.4	16.4	16.3	16.8	16.8	15.9	16.6	17.2	17	16.6	18.5	23.7	22.8	17.6	20.2	13.2
Rough	BAGUA-R	20.4	13.2	12.2	11.6	16.5	13.6	11.9	12.2	13	16.4	12	16.8	22.5	25	14.6	13.7	9.7
	CLOUD-R	18.4	14	13.6	13.5	17.4	14.7	13.6	14.6	14.3	15.9	13.4	16.8	27.4	21.7	17.1	14.6	13.3
	LION-R	30.2	19.4	17.8	17	19.4	19.1	17.9	15.8	18.4	23	16.4	25.2	20.8	29.7	15.9	15.2	13.4
	LOTUS-R	15.7	12	11.7	11.3	12.3	12.4	11	10.9	11.8	13.4	10.8	14.7	21.7	23.8	11.8	10.8	9.8
	PANDA-R	18.7	14.6	14.1	14	16	15	14.1	14.2	14.8	16.3	13.9	16.8	21.8	21.6	17.1	15.3	13.2
	M.Avg.	6.5	7.2	8.1	7.5	8.8	8.6	9.2	4.9	5.2	5.1	4.9	15	19.1	26.4	18.4	27.3	5.0
	S.Avg.	16.3	8.5	9.4	9.2	12.3	9.7	8.4	7.4	8.5	8	7.5	30.1	34.3	30.8	35.6	33.9	7.6
	T.Avg.	20.8	17	16.9	16.6	17.7	17.5	16	16.8	17.3	17.6	17	20.9	26.1	25.3	16.8	16.5	12.7
	R.Avg.	20.7	14.6	13.9	13.5	16.3	15	13.7	13.5	14.4	17	13.3	18.1	22.8	24.4	15.3	13.9	11.9
	Avg.	14.5	10.5	11	10.6	12.7	11.5	10.8	9.2	9.8	10.1	9.2	21.5	25.9	27.3	23.3	25.4	8.3

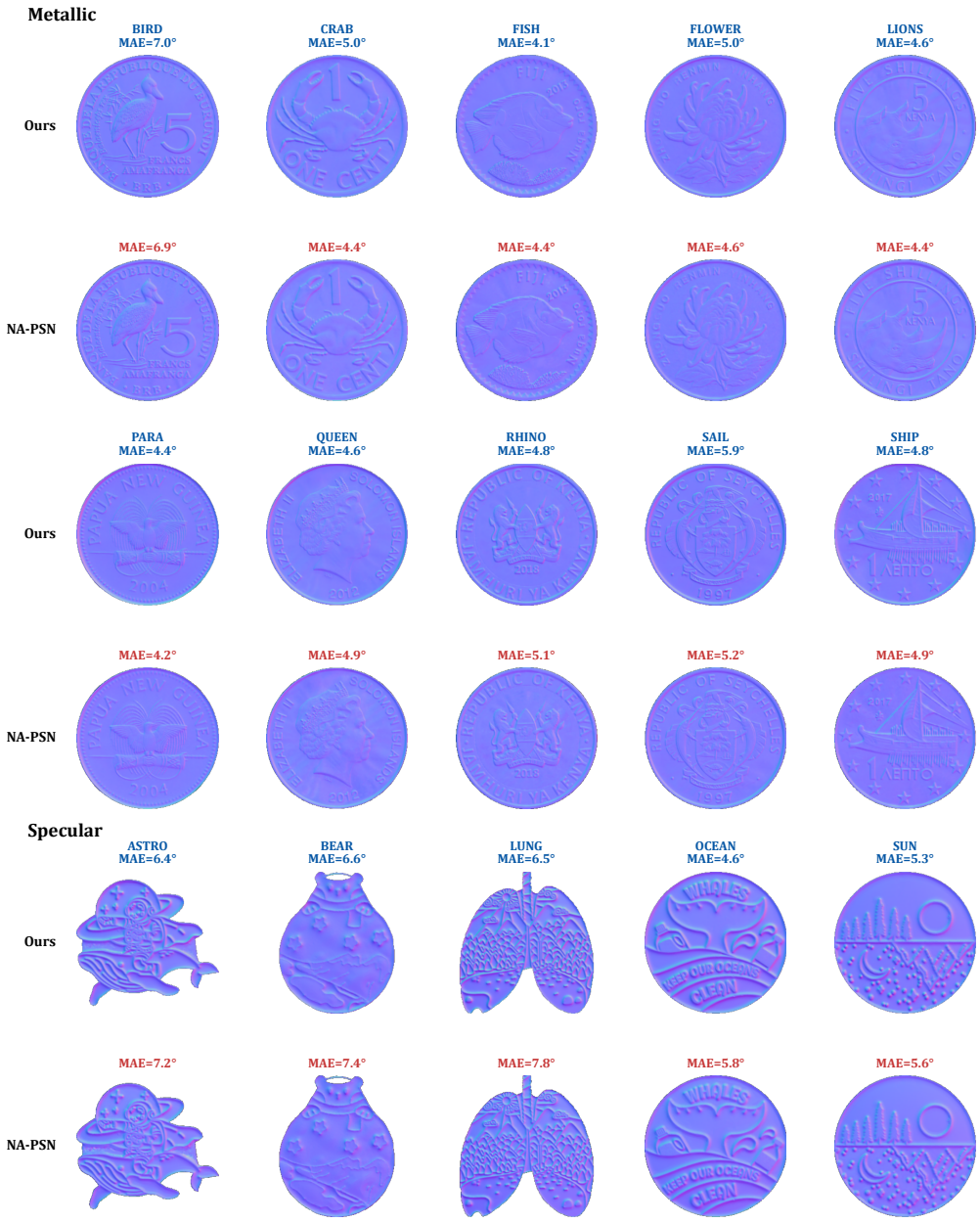


Figure 9: Comparison between our method and NA-PSN on the first 15 objects of the DiLiGenT-Pi benchmark. Each object is visualized with two rows: our prediction (top) and NA-PSN (bottom). MAE is shown below each sub-image.

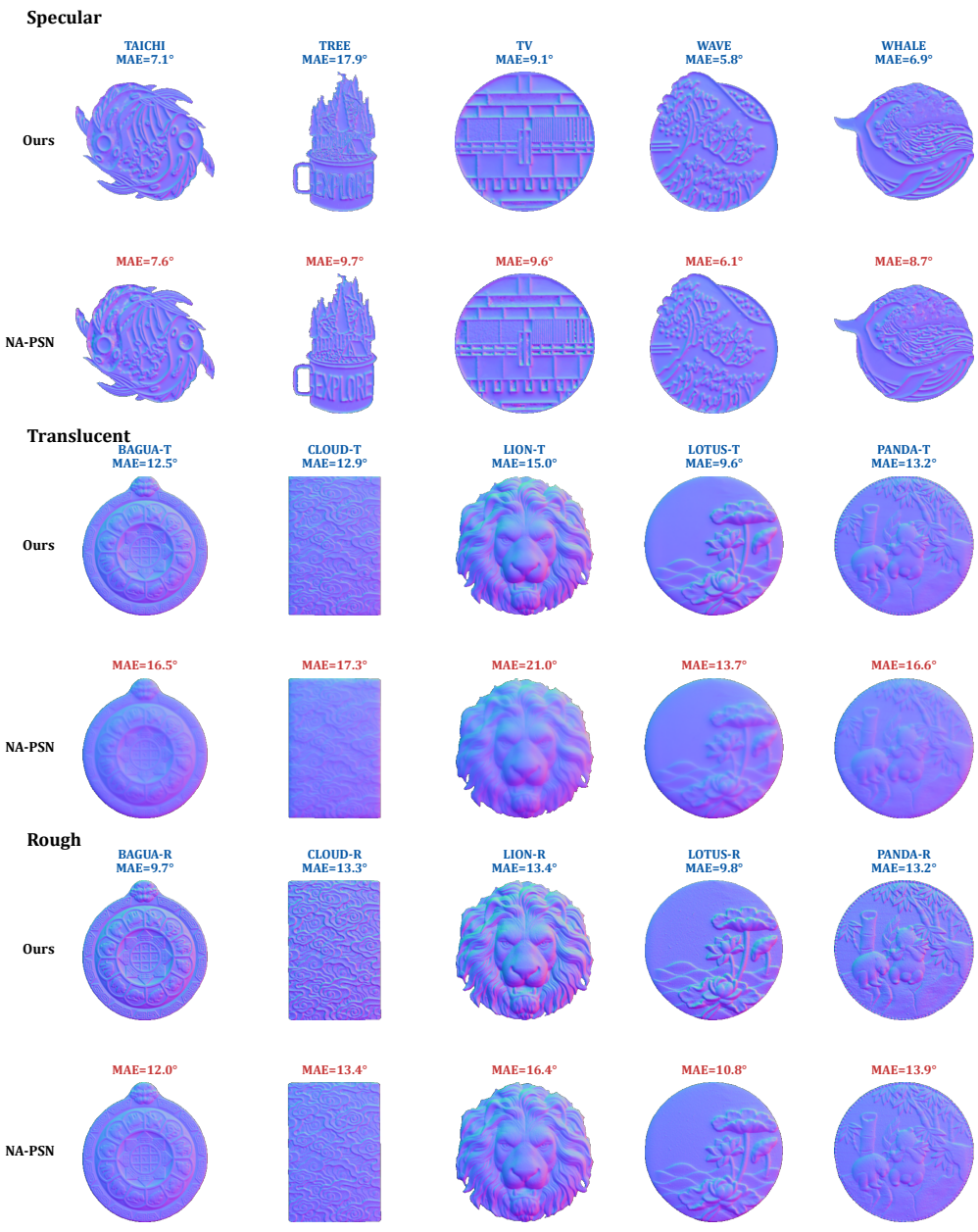


Figure 10: Comparison on the remaining 15 objects of the DiLiGenT-Pi benchmark. Category annotations (Metallic, Specular, Translucent, Rough) are displayed at the top of each group.

Table 7: Normal MAE (°) on the LUCES benchmark. LUCES contains near-field real captures with non-orthographic viewing and lens distortion, which are not included in our training assumptions. The best result for each object is highlighted in bold.

Method	Ball	Bell	Bowl	Buddha	Bunny	Cup	Die	Hippo	House	Jar	Owl	Queen	Squirrel	Tool	Avg.
HQ24	10.2	10.5	7.0	12.8	9.6	13.7	**6.2**	8.3	25.3	6.3	11.5	12.5	11.4	11.8	11.2
LC25	10.2	8.8	7.0	**12.7**	**6.1**	8.2	**6.2**	**6.0**	**22.9**	6.2	9.6	9.8	10.3	8.3	9.4
Ours	**9.5**	**8.7**	**5.4**	15.9	6.9	**5.4**	7.0	7.3	24.8	**5.0**	**7.6**	**8.6**	**9.5**	**7.4**	**9.2**

Table 8: Controlled study of principal-point offsets on N&D Syn. We report normal MAE (°) and depth δ_1.

Method	Normal MAE ↓	Depth δ_1 ↑
LC25 / GT Normal std	16.1	0.63
Baseline	5.6	0.90
0.5% shift	5.6	0.89
1% shift	5.8	0.88
3% shift	6.1	0.87
5% shift	6.5	0.84
10% shift	7.4	0.77

A.5 Failure Case Analysis & Additional Ablation

Our method is designed for uncalibrated photometric stereo under the point-light and orthographic-camera setting. Since the training data do not cover non-point-light illumination or non-orthographic imaging effects, these conditions are beyond the original scope of the model. Nevertheless, they are common in real acquisition setups, and we therefore provide additional analyses to examine the cross-domain behavior of our method.

Generalization beyond the ideal imaging setting. We first evaluate on LUCES, a real near-field benchmark with non-orthographic viewing and noticeable lens distortion. As shown in Table 7, our method achieves the best overall average normal MAE, with $9.2°$ compared with $9.4°$ for LC25 and $11.2°$ for HQ24. It also outperforms LC25 on 9 out of 14 objects. The improvement over LC25 is moderate, and we do not claim that all real-capture deviations are fully resolved. Still, the result is meaningful because the model is not trained for this setting, yet remains competitive and slightly improves over the previous state of the art on average.

Effect of principal-point offsets. Since LUCES does not provide controlled principal-point offsets, we conduct a synthetic study on N&D Syn. Starting from the baseline performance of $5.6°$ normal MAE and 0.90 depth δ_1, performance degrades only mildly under small-to-moderate shifts. When the principal point is shifted by up to 5% of the image size, the normal MAE increases from $5.6°$ to $6.5°$, while depth δ_1 remains above 0.84. A more visible drop appears only under a large 10% shift. These results suggest that the fixed-principal-point assumption is not a major issue for realistic small or moderate offsets, although very large deviations still have a noticeable effect.

Failure cases and pseudo-light reliability. On LUCES, the two objects where our method is worse than LC25 are *Buddha* and *House*. These two cases also have weaker CLIP pseudo-light statistics than the dataset average, as shown in Table 9. This suggests that the main failure factor is not the near-field setting alone, but the reliability of pseudo-light cues under complex reflectance and local geometry. Specifically, *House* mixes Lambertian and non-Lambertian regions with strong fine-scale relief, whereas *Buddha* has a mildly glossy, marble-like surface with local concavities and small specular glints. These factors make the appearance less consistent with a single illumination direction, thereby weakening pseudo-light estimation and affecting normal recovery.

A similar failure pattern is observed on the *Tree* object from DiLiGenT-Pi, the only case where our normal MAE exceeds $15°$. As shown in Fig. 11, this object contains carved, relief-like geometry with structured local bumps and ridges, together with strong non-Lambertian reflectance. The relief structures introduce localized self-shadowing and rapid normal variation, while the specular-dominant

Table 9: Failure cases on LUCES and their pseudo-light statistics. Lower normal MAE is better, while higher CLIP cosine and CLIP accuracy indicate more reliable pseudo-light cues.

Method / Metric	Buddha	House	LUCES Avg.
LC25 Normal MAE ↓	**12.7**	**22.9**	9.4
Ours Normal MAE ↓	15.9	24.8	**9.2**
CLIP Cosine ↑	0.86	0.89	0.92
CLIP Accuracy ↑	0.58	0.62	0.70

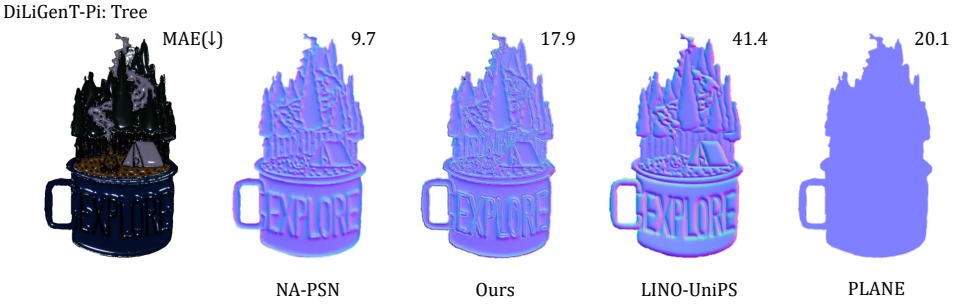


Figure 11: Failure case analysis on the *Tree* object from DiLiGenT-Pi. The object exhibits carved, relief-like geometry together with strong non-Lambertian reflectance, leading to degraded normal estimation. Even in this challenging case, our method still outperforms uncalibrated baselines.

Method	**Uncali.**	**FLOPs (G)**	**Param. (M)**	**DiLiGenT**
NA-PSN	×	481	5.5	6.5
SDM-UniPS	✓	1327	63	5.8
Our method (no CLIP)	✓	115	65	4.4
CLIP	/	51	88	/

Table 10: FLOPs and parameter comparison between our method and existing methods. DiLiGenT column shows the average normal MAE on the DiLiGenT benchmark.

reflectance disrupts the photometric consistency used by both normal estimation and depth hypothesis evaluation.

Consequently, the refined depth tends to be over-smoothed in relief regions, losing part of the local geometric variation, and the predicted normals accumulate larger errors around strongly specular areas. This combination of relief-like geometry and strong non-Lambertian reflectance represents a rare and extreme corner case in the benchmark. For typical non-Lambertian objects or common high-frequency shapes, the method remains stable; even on *Tree*, it still substantially outperforms uncalibrated baselines, confirming the robustness of the overall pipeline in practical scenarios.

A.6 Computational Complexity and Parameter Comparison

Our method consists of two main steps: CLIP-based calibration pseudo-lighting and model inference. The total FLOPs consumed in our approach primarily come from the part where CLIP uses a ViT model to map images into vectors (since the text is fixed, no additional computation is required during the evaluation process for the text, and thus, only the FLOPs from the ViT are considered). We adopt the ViT-B/32 model. In the case of 12 input images of size 224x224, the comparison with existing methods is presented in Table. 10.

The FLOPs for CLIP in the table account for the total inference required for the 12 images. As shown, while our method involves more modules, it does not significantly increase the computational burden. Though our parameter count is higher compared to the calibrated method, this is typical for Uncalibrated photometric stereo methods. The total parameter count of our method is not much larger than that of other Uncalibrated methods, even considering the additional parameters introduced by CLIP, and it is still much smaller than existing high-precision Uncalibrated methods.

A.7 AGI USAGE STATEMENT

In this work, we made limited use of AGI tools (GPT-5.0/5.1 and Qwen3-Max) strictly for grammar refinement and expression simplification during manuscript preparation, based on content originally written by the authors. A small amount of non-essential visualization code was also assisted by AGI, while all core algorithms, model implementations, and experimental pipelines were independently developed by the authors. AGI did not contribute to method design, scientific reasoning, or experimental conclusions; all technical contributions originate solely from the authors.