COIL-PS: Continuous and Online Illumination Planning for Photometric Stereo

JUN HOONG CHAN,[1,2] BOHAN YU,[1,2,*,†] JIEJI REN,[4] JINXIU LIANG,[3,*,†] AND BOXIN SHI[1,2]

[1]*State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University, Beijing 100871, China.*
[2]*National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China.*
[3]*National Institute of Informatics, Tokyo, 101-8430, Japan.*
[4]*School of Mechanical Engineering Shanghai Jiao Tong University, Shanghai 200240, China.* [†]*These authors are co-corresponding authors.*
[*]*ybh1998@pku.edu.cn, cssherryliang@gmail.com*

Abstract: Photometric Stereo (PS) aims to recover high-fidelity surface normals by observing pixel-wise radiometric variations under different light directions. However, traditional PS methods require dense sampling of the incident light to mitigate non-Lambertian effects, such as cast shadows and specular highlights, creating a significant efficiency bottleneck for practical optical metrology. To address this efficiency bottleneck, illumination planning methods seek to identify an optimal set of light directions to maximize information gain with minimal measurements. A critical limitation of existing illumination planning paradigms is their reliance on selecting from a predefined, discrete set of candidate light directions. This discretization of the light space introduces an artificial bottleneck, severely limiting precision and adaptability. In this paper, we address this limitation by introducing a **C**ontinuous and **O**nline **IL**lumination Planning framework for **P**hotometric **S**tereo **(COIL-PS)**. Instead of selecting from a fixed grid, our method formulates illumination planning as a continuous regression problem, adaptively steering light positioning in the continuous hemispherical domain. By coupling online illumination planning with feedback from intermediate normal estimates, COIL-PS adaptively navigates non-Lambertian effects such as shadows and specularities, which allow for the precise angular placement of illumination required to resolve geometric ambiguities that fall between fixed grid points. Extensive experiments on a synthetic dataset, semi-real benchmarks, and our custom-built real-world robotic validation system demonstrate that COIL-PS achieves superior normal reconstruction accuracy compared to state-of-the-art discrete planning methods, even with a budget of only ten lights, significantly outperforming discrete planning paradigms.

1. Introduction

The precise recovery of 3D surface orientation is a fundamental objective in optical metrology, industrial inspection, and cultural heritage digitalization [1–4]. Photometric Stereo (PS), first formalized by Woodham [5], is used for surface recovery by inverting the image formation model: observing how the radiance of a static surface changes under different light directions with a fixed camera viewpoint. For an ideal Lambertian surface, a minimum of three linearly independent light sources suffices to determine the surface normal. However, real-world materials exhibit complex reflectance properties—including specular highlights, cast shadows, and global illumination effects like inter-reflections—fundamentally violate the linear Lambertian model. Modern PS techniques, ranging from classical robust estimation [6–8] to data-driven deep learning methods [9–14], have made significant strides in handling these effects. However, they typically require dense sampling, capturing 50 or more images to robustly filter out outliers [15]. This dense capture requirement creates a critical bottleneck for real-world adoption: in scenarios ranging from industrial inspection to robotic manipulation [1–4], acquisition time is a scarce

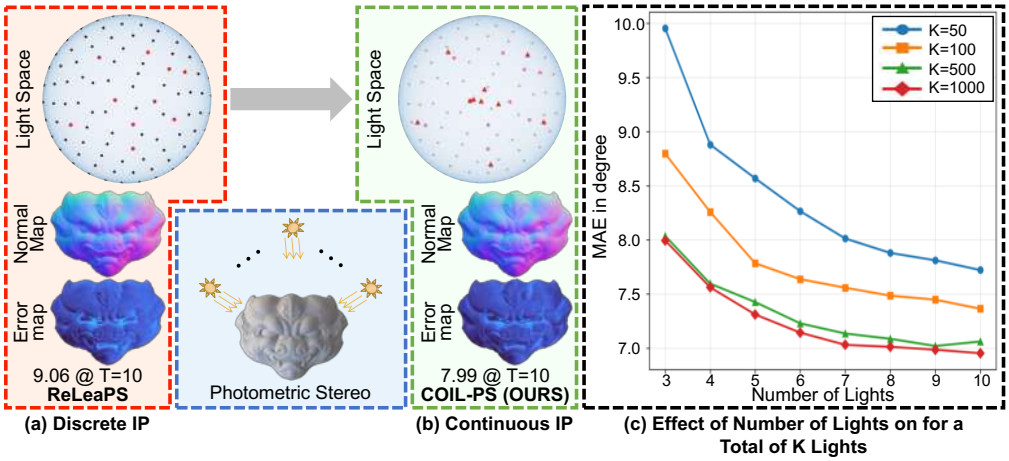


Fig. 1. The advantage of continuous illumination planning (IP) for photometric stereo (PS) with a budget of $T = 10$ (red dots) on the "Monster" object. (a) **Discrete IP:** ReLeaPS [18] is constrained to select light directions from 100 fixed candidates (black dots). Even if the ideal lighting angle lies between grid points, the method is forced to choose a suboptimal neighbor, resulting in quantization error. (b) **Continuous IP:** Our COIL-PS predicts a set of $T = 10$ light directions (red triangles) as continuous vectors on the hemisphere. Notably, these predicted directions frequently lie between the predefined discrete lights (gray dots). This is in contrast to the light directions estimated by ReLeaPS [18] (pink dots), highlighting the difference between discrete and continuous IP. This flexibility allows the system to place lights precisely where needed to resolve geometric ambiguities, leading to sharper details and lower error using the same budget. (c) **The limitation of discretization:** MAE for exhaustive (brute-force) selection on the Sculpture dataset [21], varying the number of captured lights T from 3 to 10; curves correspond to candidate set sizes $K \in \{50, 100, 500, 1000\}$. The results reveal that increasing K (making the grid denser) monotonically lowers the error, validating that lifting the discrete constraint is essential for maximizing performance.

resource, objects may be dynamic, and deploying complex multi-light systems is often impractical.

This tension between reconstruction fidelity and acquisition efficiency has motivated research in Illumination Planning (IP) [16–20]. The fundamental objective of IP is to maximize the information from each measurement, shifting the paradigm from capturing *more* data to capturing *better* data. The central question becomes: given a strict budget of T lights (*e.g.*, $T = 10$), which specific light directions maximize the conditioning of the inverse problem for normal recovery? An effective IP strategy must identify illumination configurations that minimize degenerate shading conditions (such as attached shadows) while maximizing photometric diversity across the surface. Existing approaches have tackled this problem through either the *offline* computation of universal lighting configurations [16, 17, 20] or learned *online* policies that adapt to specific objects [18, 19]. However, nearly all previous work shares a common fundamental limitation: it formulates the problem in a **discretized light space**. These methods restrict the search space to a fixed candidate set of K available lights (*e.g.*, $K = 100$) and reduce the optimization task to a combinatorial selection of a subset of T from K candidates. As illustrated in Fig. 1 (a), this *discrete* formulation forces the selection of suboptimal light vectors when the ideal sampling angle falls between grid points. This discretization is an algorithmic convenience that contradicts the physical reality in which light can arrive from any direction within the continuous hemisphere.

Why is this discretization problematic? The empirical analysis in Fig. 1 (c) validate the theoretical expectation: as we increase the density K of the discrete set to approximate a continuous hemisphere, the Mean Angular Error (MAE) progressively decreases. Denser sampling—closer to the continuous ideal—yields a more accurate normal measurement. This occurs because optical

features such as shadow boundaries and specular lobes are high frequency angular functions; the optimal illumination angle required to resolve a specific geometric ambiguity often lies between predefined grid points. Consequently, previous methods inherently limit the system's ability to probe the surface reflectance functions precisely where they are most informative.

Transitioning from a discrete selection task (essentially a classification problem) to continuous coordinate prediction (a regression problem) presents technical challenges. The mapping between a continuous light direction and the reconstruction error is highly non-convex and discontinuous. A minor perturbation in the light vector can cause a pixel to abruptly transition from illuminated to shadow or trigger a saturated specular spike, resulting in discontinuities in the loss landscape. Successfully navigating this complex manifold requires a system capable of modeling the temporal evolution of the surface appearance and reasoning about geometric uncertainty.

In this paper, we introduce **C**ontinuous **O**nline **IL**lumination Planning for **P**hotometric **S**tereo **(COIL-PS)**. We formulate IP as a sequential decision-making process in the continuous domain. Fig. 1 (b) illustrates our *continuous* setting, where COIL-PS predicts $T = 10$ optimal light vectors as continuous coordinates on the hemisphere, enabling precise placements of illumination sources between standard discrete grid points. Our method employs a closed-loop architecture driven by the residual uncertainty from the current estimation of normals. It integrates a temporal sequence processor with a causal attention mechanism [22] to aggregate photometric history and a learnable outlier-aware gating mechanism that acts as a dynamic validity mask to identify and suppress radiometric data corrupted by non-Lambertian effects. By coupling this continuous, feedback-driven policy with PS normal estimation, COIL-PS effectively steers angular configurations that maximize the resolution power of the reconstruction. Our contributions are:

- To our knowledge, we are the first to formulate and solve the IP problem for PS directly in the continuous hemispherical domain, addressing the quantization limitations inherent in discrete candidate sets.
- We propose an online, feedback-driven adaptive sampling policy network that incorporates a causal attention mechanism for the temporal integration of radiometric signals and a dynamic outlier-aware gating mechanism to robustly filter optical outliers, such as specularities and shadows.
- We provide extensive validation on synthetic and semi-real benchmarks, demonstrating state-of-the-art reconstruction accuracy. Furthermore, we validate the practicality of our method on a custom-built real-world robotic validation system, showing that our continuous policy can effectively drive physical hardware to achieve superior metrological results.

The remainder of this paper is organized as follows. Section 2 reviews related work on PS and IP. Section 3 introduces the methodology of the proposed COIL-PS framework. Section 4 describes the experimental setup and results. Finally, Section 5 concludes the paper.

2. Related Work

In this section, we review prior work on PS and IP, highlighting advances in robust and learning-based normal estimation methods, as well as offline and online strategies for optimizing lighting configurations.

2.1. Photometric Stereo

Conventional PS [5] employs least-squares estimation with a minimum of three images captured under distinct light directions to recover surface normals for Lambertian objects. To handle general reflectance and non-Lambertian surfaces, prior work falls into two paradigms. The **classical paradigm** uses robust formulations to mitigate shadows and specularities [6–8, 23–26], operating on the principle that accurate normals can be recovered if sufficient observations are

available to "vote out" outliers. These methods typically require 10–30 images. The **learning-based paradigm** approximates the complex mapping from images to surface normals using deep neural networks. These include pixel-wise models (*e.g.*, DPSN [9], CNN-PS [10], Attention-PSN [11], PX-Net [27]) and image-based models that leverage spatial context (*e.g.*, PS-FCN [12], SDPS-Net [13], IRPS [14]). Learning-based methods are typically trained and evaluated with dense batches of approximately 50–100 images. While these models generalize well to complex reflectance given sufficient illumination diversity, such dense capture requirements are often impractical for real-world deployment due to object motion, limited control over lighting, and tight acquisition budgets. Motivated by this efficiency gap, we pursue continuous online IP to maximize reconstruction accuracy under a sparse lighting budget, effectively balancing data efficiency with metrological fidelity.

2.2. Illumination Planning in Photometric Stereo

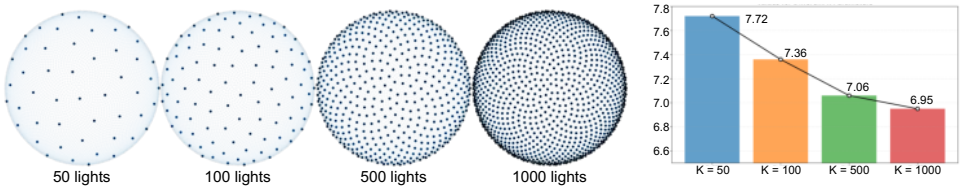
Instead of passively processing a fixed dataset, IP for PS actively selects the optimal lighting configuration to maximize information gain. Existing IP methods can be broadly categorized into *offline* and *online* strategies. **Offline IP methods** compute universal lighting configurations that are independent of the specific object being scanned. DC05 [16] introduces one of the earliest formulations for Lambertian PS by modeling camera noise, demonstrating that any orthogonal triplet of light directions is optimal regardless of geometry. While foundational, this shape-agnostic theory fails to account for cast shadows or specularities that depend on specific object geometry. Building on this, TK22 [17] proposes an iterative offline strategy that reduces shadow effects while remaining shape-agnostic. More recently, LIPIDS [20] presents a learning-based offline approach that learns universal lighting configurations transferable across diverse scenes. However, these offline methods cannot adapt to the unique geometry of the object being scanned, often wasting the lighting budget on redundant or uninformative angles. **Online IP methods** adapt lighting to the specific object during acquisition. ReLeaPS [18] and LAC-PS [19] employ reinforcement learning to sequentially select light directions through reward-driven policies. ReLeaPS [18] optimizes lighting for generalized PS under limited illumination, whereas LAC-PS [19] augments this with an accuracy-assessment network that predicts reconstruction quality without ground-truth normals. Despite these advances, *all existing IP methods rely on a discrete, predefined set of candidate light directions*. This discretization creates a quantization gap, preventing the system from probing the reflectance function at the precise angular coordinates required to resolve fine-grained geometric ambiguities. These limitations motivate our continuous illumination-planning formulation, which optimizes planning directly in the full hemispherical domain, enabling more flexible and physically faithful adaptation.

3. Continuous Online Illumination Planning

In this section, we detail the proposed COIL-PS framework. We begin with the definition of the PS objective and the limitations of discrete IP, followed by the formulation of our continuous IP framework. We then present the reconstruction-driven optimization strategy used to train the framework. Finally, we describe the adaptive sampling architecture for the joint optimization of IP and normal estimation, focusing on the modules designed to handle non-Lambertian radiometric outliers and integrate temporal measurement history.

3.1. Problem Formulation

Photometric Stereo. PS reconstructs surface geometry by inverting the image formation model that connects surface orientation to observed irradiance. Assuming an orthographic camera with a linear radiometric response and T calibrated directional lights, the i-th image $I_i \in \mathbb{R}^{H\times W}$ is captured under light direction $l_i \in \mathbb{R}^3$, where H and W denote image height and width. For a



(a) Fibonacci-sampled Light Distribution **(b) Effect of Light Density K on MAE at T=10**

Fig. 2. Empirical motivation for continuous planning. (a) Visualization of Fibonacci-sampled light directions with increasing density K. Discrete IP methods are mathematically restricted to these fixed lattices. (b) MAE for an exhaustive brute-force search for the optimal 10-light configuration on the Sculpture dataset. The MAE decays asymptotically as K increases (7.72° at $K = 50 \rightarrow 6.95°$ at $K = 1000$) as the grid become denser.

pixel $p \in \Omega$, where Ω denotes the set of valid pixels, the image intensity $I(p)$ under l is given by:

$$I(p) = \mathcal{F}\big(n(p), l, v\big) \max(l \cdot n(p), 0) + \epsilon(p), \tag{1}$$

where $n(p) \subset \mathbb{S}^2_+ \in \mathbb{R}^3$ is the unit surface normal, v is the fixed view direction under orthographic projection, $\mathcal{F}$ is the reflectance model, and $\epsilon(p)$ accounts for sensor noise and unmodeled effects, such as residual global-illumination effects. The max operator handles attached shadows by truncating negative dot products.

The objective of PS is to estimate a surface normal map that minimizes the discrepancy between the observed irradiance and the physical model by inverting this image formation process. We quantify reconstruction accuracy using the MAE between the estimated and ground-truth surface normal maps:

$$\text{MAE} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \arccos\big(n(p) \cdot n_{\text{gt}}(p)\big). \tag{2}$$

Given a stack of T observed images $\mathcal{I} = \{I_1, I_2, \ldots, I_T\} \in \mathbb{R}^{H \times W \times T}$ captured under known light directions $L = \{l_1, l_2, \ldots, l_T\} \in \mathbb{R}^{T \times 3}$, photometric stereo tasks seek an algorithm $\text{PS}(\cdot)$ that produces an estimated surface normal map $N \in \mathbb{R}^{H \times W \times 3}$:

$$N = \text{PS}(\mathcal{I}, L) \tag{3}$$

that minimizes the MAE in Eq. (2).

Continuous Illumination Planning. In PS, IP selects T light directions that minimize the expected reconstruction error. The conventional approach chooses from a predefined set of candidate light directions $L_{\text{discrete}} = \{l_1, \ldots, l_K\}$ on a discretized hemisphere. This problem, known as *discrete illumination planning*, restricts the solution space to a finite set of predefined light directions. We posit that this discretization imposes a quantization gap on performance.

To quantify this gap, we evaluate Fibonacci-sampled light directions [28] with increasing K on the Sculpture dataset [21] using a brute-force selection strategy. For each of the $T = 10$ steps, a single-step optimal light direction is determined via an exhaustive, brute-force search over all sampled light directions, selecting the one that yields the lowest MAE through a least-squares method [5]. As shown in Fig. 2, the MAE decreases monotonically as K increases. This indicates that lifting the discrete constraint is essential for maximizing performance.

Motivated by this observation, we formulate IP directly on the unit hemisphere $\mathbb{S}^2_+$, optimizing light directions as continuous vectors rather than choosing from a fixed candidate set L_{discrete}. The process begins by capturing an initial image I_1 under a predefined light direction l_1. For each step $i \in \{1, \ldots, T-1\}$, the model uses the current image stack $\mathcal{I}_i = \{I_j\}_{j=1}^{i}$ and the corresponding light directions $L_i = \{l_j\}_{j=1}^{i}$ to predict the next direction $l_{i+1} \in \mathbb{S}^2_+$. We then acquire the image

I_{i+1} under l_{i+1} and update the normal estimation as $N_{i+1} = \mathrm{PS}(\mathcal{I}_{i+1}, L_{i+1})$, which is used to optimize Eq. (4). The new observation is appended to the stack, and the process repeats until T light-image pairs are acquired. We refer to this as *continuous illumination planning*:

$$L_\star^{(\mathrm{cont})} = \underset{\substack{L_T \subseteq \mathbb{S}_+^{T\times 2} \\ |L_T|=T}}{\arg\min} \frac{1}{HW} \left[1 - (\mathrm{PS}(\mathcal{I}_T, L_T) \cdot N_{\mathrm{gt}})\right]. \tag{4}$$

This formulation extends IP from a finite candidate lattice to a continuous domain, allowing the model to explore the full hemispherical space for precise light placement. In practice, the model predicts unconstrained Cartesian vectors $l_i = (x_i, y_i, z_i) \in \mathbb{R}^3$ and maps them to $l_i \in \mathbb{S}_+^2$ by enforcing $\|l_i\|_2 = 1$ and $z_i \geq 0$. This eliminates any dependence on predefined discrete candidate light directions, effectively simulating $K \to \infty$.

3.2. Reconstruction-Driven Optimization

Since the "optimal" light sequence depends entirely on the unknown geometry, we cannot supervise the light directions directly. Instead, we train COIL-PS end-to-end using reconstruction accuracy as the supervisory signal and introduce a reconstruction-driven normal loss that couples light selection with normal reconstruction by directly optimizing the PS objective. Note that MAE (Eq. (2)) is used for evaluation, whereas training relies on a cosine-similarity loss.

At each planning step i, after acquiring image I_{i+1} under light l_{i+1}, we update the input stacks to include $\mathcal{I}_{i+1}$ and L_{i+1}. The PS backbone then produces a normal estimate $N_{i+1} = \mathrm{PS}(\mathcal{I}_{i+1}, L_{i+1})$, yielding the normal vector $n_{i+1}(p)$ for each pixel p. Supervision is applied via the cosine-similarity loss between the estimated and ground-truth normals:

$$\mathcal{L}_{i+1} = \frac{1}{|\Omega|} \sum_{p\in\Omega} \Big(1 - n_{i+1}(p) \cdot n_{\mathrm{gt}}(p)\Big). \tag{5}$$

To prioritize later timesteps that progressively accumulate more informative observations, we compute the aggregate loss with exponentially increasing weights:

$$\mathcal{L}_{\mathrm{normal}} = \sum_{i=i_{\min}}^{T} w_i \, \mathcal{L}_{i+1}, \qquad w_i = 1 - 0.9^{\,i}, \tag{6}$$

where we set $i_{\min} = 3$ to avoid degenerate supervision when too few lights are available.

A key challenge in training is to ensure that the image formation process is differentiable with respect to light directions. The previous learning based PS methods [9–14, 27] are usually trained on synthetic datasets generated from physics-based ray-tracing renderers [29–31]. These synthetic datasets are discrete and non-differentiable. There are also some differentiable rendering methods [32, 33] to make the image formation model differentiable, but they are too time-consuming for our training process. To enable efficient and accurate gradient back-propagation, we introduce a differentiable soft-interpolation strategy during training. Given a predicted light direction l_{i+1} and the set of all available light directions $\{l_p\}_{p=1}^K$, we compute the angular distances θ_p and convert them into interpolation weights w_p using a softmax function with a temperature parameter τ:

$$\theta_p = \arccos\big(l_{i+1} \cdot l_p\big), \qquad w_p = \frac{\exp(-\theta_p/\tau)}{\sum_{p=1}^{T} \exp(-\theta_p/\tau)}. \tag{7}$$

The temperature τ modulates the sharpness of the weight distribution, we choose a small value of τ to make the weights more concentrated toward the nearest light directions. Using these soft weights w_p, an interpolated image and its corresponding light direction are computed as follows:

$$\tilde{I} = \sum_{p=1}^{T} w_p I_p, \qquad \tilde{l} = \sum_{p=1}^{T} w_p l_p. \tag{8}$$

This interpolation yields a continuous approximation of the image formation process for the continuous prediction of light direction while maintaining a differentiable path from the reconstruction loss back to the light selection network. Consequently, the weighted-aggregated normal loss in Eq. 6 backpropagates through the interpolated image $\tilde{I}$, the normal estimation $\mathrm{PS}(\cdot)$ method, and the light direction selection network (discussed in the following subsection), enabling end-to-end optimization and providing smooth gradients that guide light selection toward illumination conditions in continuous light space that most effectively reduce the MAE.

3.3. Light Direction Selection Network

Overview. We propose COIL-PS, a continuous online IP network that predicts the optimal light direction in the continuous hemispherical domain, conditioned on the current image observations and light direction history. The illumination planning policy must be robust to non-Lambertian effects that necessitate multi-capture. We propose a recurrent architecture to fuse visual appearance from the inputs with an encoded sequence of historical light directions. As shown in Fig. 3, COIL-PS comprises three core modules: (a) a *Multi-Scale Image Encoder* with an *Outlier-Aware Gating* mechanism, (b) a *Temporal Sequence Processor with Causal Attention*, and (c) an *Appearance-Light Fuser* with a *Light Direction Decoder*. A key component is the Outlier-Aware Gating mechanism, which dynamically assigns per-pixel reliability to suppress noisy regions corrupted by non-Lambertian effects.

Multi-Scale Image Encoder with Outlier-Aware Gating Mechanism. Real-world images are often corrupted by photometric outliers (*e.g.*, deep shadows and saturated specularities) that violate PS assumptions. To handle this, we employ a multi-scale CNN encoder that processes both input images and a dynamically predicted outlier mask. Given the current stack of i images $\mathbf{I} \in \mathbb{R}^{B\times i\times H\times W\times 3}$ (RGB), we flatten the batch and sequence dimensions to $\tilde{\mathbf{I}} \in \mathbb{R}^{(Bi)\times H\times W\times 3}$ and feed them to the *Multi-Scale Image Encoder*. The encoder comprises eight weight-normalized convolutional layers with ReLU activations that expand the per-frame channel depth from $3 \to 32 \to 64 \to 128 \to 256$, interleaved with 2×2 max-pooling layers to progressively halve the spatial resolution to 8×8. After encoding, we restore the sequence dimension to obtain per-frame feature maps $\mathbf{F}_i \in \mathbb{R}^{8\times 8\times 256}$. Stacking over batch and sequence yields $\mathbf{F}_{\text{img}} \in \mathbb{R}^{B\times i\times 8\times 8\times 256}$, providing 256-dimensional features at each of the 64 spatial locations.

While the encoder aggregates informative cues, outliers such as deep shadows and interreflections corrupt features and degrade downstream estimation. A common practice [34] is to apply hard intensity clipping; however, such rules are non-differentiable and incompatible with end-to-end learning. Moreover, a fixed threshold cannot adapt to material- and illumination-dependent variability across frames and scales. To address this, we introduce the *Outlier-Aware Gating* mechanism: a lightweight mask generated based on the input images, which is used to predict per-pixel reliability from normalized intensities via a temperature-controlled sigmoid with two learnable thresholds. The mask branch shares the same *Multi-Scale Image Encoder* as the image stream, and per-frame outlier masks $\mathbf{M} \in (0,1)^{B\times i\times H\times W}$ are flattened to $\tilde{\mathbf{M}} \in \mathbb{R}^{(Bi)\times H\times W}$ and passed through the encoder, yielding multi-scale mask embeddings $\mathbf{F}_i^M \in \mathbb{R}^{8\times 8\times 256}$. The two learnable thresholds, θ_{low} and θ_{high}, determine what the network considers to be a shadow or a highlight, and their values are automatically adjusted to best suppress those unwanted regions. Masks are recomputed online at every step i and remain differentiable, ensuring that the gradients from the reconstruction loss jointly update the gating parameters. Restoring the batch and sequence dimensions produces $\mathbf{F}_{\text{mask}} \in \mathbb{R}^{B\times i\times 8\times 8\times 256}$, which modulates the corresponding image features $\mathbf{F}_{\text{img}}$ through element-wise gating $\tilde{\mathbf{F}} = \mathbf{F}_{\text{mask}} \odot \mathbf{F}_{\text{img}}$, yielding masked feature tensors.

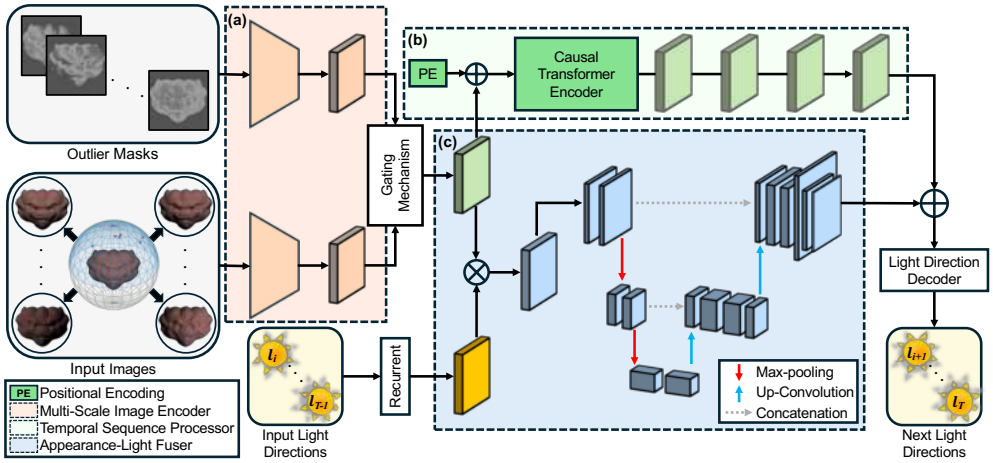


Fig. 3. The Architecture of the COIL-PS network, which operates as a recurrent loop. (a) *Multi-Scale Image Encoder*: Extracts visual features from the input stack. An *Outlier-Aware Gating* mechanism generates a reliability mask to suppress non-Lambertian artifacts (shadows/specularities) before feature aggregation. (b) *Temporal Sequence Processor*: A transformer encoder with a causal attention mask integrates the photometric history for each pixel, modeling the evolution of appearance over time. (c) *Appearance-Light Fuser and Light Direction Decoder*: Combines temporal visual features with spatially expanded embeddings of the lighting history. Finally, the *Light Direction Decoder* regresses predicts the next optimal continuous light direction.

Temporal Sequence Processor with Causal Attention. The sequence of observations matters; to predict the next light direction, the model must condition on the observation history. We employ a transformer encoder [22] that operates independently at each spatial location, ingesting the time-ordered features observed at that pixel up to the current step. This *Temporal Sequence Processor* models the temporal evolution of encoded image features, capturing illumination-dependent appearance changes across the input sequence and using that history—causally, without access to future frames—to inform the next-light prediction.

It operates on the masked feature tensor $\tilde{\mathbf{F}} \in \mathbb{R}^{B\times i\times 8\times 8\times 256}$, where each spatial location from the 8×8 grids forms an independent temporal sequence describing the photometric evolution of that pixel across i input frames. For each spatial location, the 256-dimensional features are projected into a higher-dimensional latent space (d=1024), while Positional Encoding (PE) [22] injects frame-order information. This per-pixel temporal modeling enables the network to learn illumination-consistent reflectance cues under dynamically changing lighting. Temporal reasoning is performed by a transformer encoder [22] equipped with a causal attention mask, ensuring that each feature depends only on current and preceding observations—thus supporting autoregressive online prediction. The refined embeddings are projected back to the original feature dimensionality and reshaped to (8×8×256) per frame, yielding temporally enriched features $\tilde{\mathbf{F}}_{\text{img}} \in \mathbb{R}^{B\times i\times 8\times 8\times 256}$. A residual connection with the input preserves spatial detail while integrating sequential context. A series of weight-normalized convolutional layers further refines the features, progressively condensing the channel dimensionality from 256 to 1 while preserving the 8×8 spatial resolution, resulting in $\tilde{\mathbf{F}}_{\text{v}} \in \mathbb{R}^{B\times i\times 8\times 8}$.

Appearance-Light Fuser. The next-light decision must condition on both the appearance captured by temporally processed features and the lighting history (*i.e.*,the actions taken so far). To capture this dependency, we introduce the *Appearance-Light Fuser*, which integrates the temporally refined visual representation with an encoded history of light directions to form a joint state that guides subsequent light selection. It receives the masked and temporally processed image features $\tilde{\mathbf{F}} \in \mathbb{R}^{B\times i\times 8\times 8\times 256}$ and the corresponding light directions $L_i \in \mathbb{R}^{B\times i\times 3}$ up to

timestep i. Each light direction is projected through a light embedding layer composed of two fully connected transformations (3→32→64) with ReLU activations, producing $L_{\text{enc}} \in \mathbb{R}^{B\times i\times 64}$. The encoded light vectors are spatially expanded to match the spatial extent of the image features, yielding light-aware feature maps $L_{\text{map}} \in \mathbb{R}^{B\times i\times 8\times 8\times 64}$. These are concatenated with the first 64 channels of the image features, forming fused appearance–illumination representations $\tilde{\mathbf{F}}_{\text{fuse}} \in \mathbb{R}^{B\times i\times 8\times 8\times 128}$.

The fused tensor is processed through a U-Net–style encoder–decoder network consisting of weight-normalized 3×3 convolutional and transposed-convolutional layers that progressively encode and reconstruct multi-scale contextual features. The encoder gradually increases channel depth (128→256→512) while downsampling the spatial resolution via pooling, and the decoder symmetrically upsamples and concatenates intermediate features to recover fine-grained spatial details. Each stage employs ReLU activations and skip connections to preserve local structure and stabilize training. The final convolutional layers condense the channel depth from 128 to 1, producing a fused output $\tilde{\mathbf{F}}_{\text{a}} \in \mathbb{R}^{B\times i\times 8\times 8}$. This output encapsulates both spatially localized appearance cues and temporally conditioned lighting contexts, serving as input to the *Light Direction Decoder*.

The *Light Direction Decoder* translates the fused representations into an explicit prediction of the optimal light direction. It takes as input the temporally processed visual features $\tilde{\mathbf{F}}_{\text{v}}$ and the fused appearance–illumination features $\tilde{\mathbf{F}}_{\text{a}}$, which are combined through a residual operation $x = \tilde{\mathbf{F}}_{\text{v}} + (\tilde{\mathbf{F}}_{\text{a}} - \tilde{\mathbf{F}}_{\text{a}}^{\text{mean}})$ to emphasize illumination-dependent variations while preserving scene structure. The fused tensor is flattened and passed through a lightweight fully connected predictor (64→32→3) with ReLU activation to produce light vectors $L_{i+1} \in \mathbb{R}^{B\times\{i+1\}\times 3}$. The x– and y–components correspond to the azimuthal plane, while the z–component is constrained to be positive through absolute rectification and is stabilized with a small constant (10^{-8}). The predicted light directions are normalized to unit length, ensuring physically valid orientations confined to the upper hemisphere. This prediction is a sequence of light directions, and the next light direction l_{i+1} is always taken from the final prediction in this sequence, representing the online decision based on all preceding observations. This decoder completes the end-to-end COIL-PS pipeline by mapping the fused spatiotemporal features to a continuous illumination vector that guides the selection of the next light direction.

4. Experiments

We conduct a comprehensive evaluation of the proposed COIL-PS against the previous state-of-the-art IP methods. We test on a synthetic dataset, semi-real benchmarks, and our custom-built real-world robotic validation system. Quantitative and qualitative analyzes demonstrate that COIL-PS achieves superior reconstruction accuracy, highlighting the practical advantages of continuous illumination planning.

Compared Methods. Our experiments compare COIL-PS with both offline and online IP methods, including random light selection, DC05 [16], TK22 [17], ReLeaPS [18], LIPIDS [20], and LAC-PS [19]. All methods are given the same budget of $T = 10$ lights and predict the surface normal using least-squares estimation [5].

Implementation Details. We train on synthetic datasets for 500 epochs using the AdamW optimizer (learning rate 1×10^{-5}, weight decay 1×10^{-8}), a constant learning rate schedule, a batch size of 8, and planning steps of $T = 10$. Before evaluating the cosine term, both predicted and ground-truth normals are ℓ_2-normalized. Model selection is based on the lowest validation MAE, and all reported results are generated from the selected checkpoint using least-squares estimation during the validation stage. Training on a single NVIDIA GeForce RTX 3090 takes approximately 15 hours.

4.1. Datasets

We train on synthetic images rendered from the Sculpture [21] shape collections and evaluate the generalization performance on the DiLiGenT [15] and DiLiGenT10^2 [35] semi-real benchmarks, as well as on images captured with our custom-built real-world robotic validation system for the "Monster" and "Sphere" objects. For evaluation on the semi-real benchmarks and synthetic datasets, we also utilize the differentiable soft-interpolation strategy (Eq. 8) to approximate the images under continuous illumination conditions.

Synthetic Datasets. To simulate a pseudo-continuous light space, we generate synthetic data using a physics-based ray-tracing renderer [33]. The renderer produces synthetic images under 1000 distinct light directions, sampled via the Fibonacci technique [28] for shapes from the Sculpture [21] dataset. The dataset is augmented with random scaling, rotation, and translation, yielding 1000 training samples and 100 test samples. We further employ a soft-weight interpolation module to smoothly approximate continuous illumination from the discretely sampled light directions.

Semi-Real Benchmark. For evaluation in real-world conditions, we use the DiLiGenT [15] benchmark. This benchmark comprises 10 objects exhibiting a wide range of shapes and surface materials, with each object captured under 96 distinct light directions and accompanied by ground-truth normal maps. This setup facilitates a rigorous assessment of the generalization ability of the proposed method from synthetic training data to real-world PS. In addition, we also include the DiLiGenT10^2 [35] benchmark, which provides higher-resolution measurements and an expanded set of lighting conditions. Incorporating DiLiGenT10^2 [35] allows us to further evaluate the robustness of the proposed method under more challenging real-world settings.

Hardware Verification Targets. We utilize two physical target objects—a "Monster"with complex non-convex geometry and a "Sphere" with ideal convex geometry—both 3D-printed directly from their corresponding CAD meshes. These targets are integrated into our real-world robotic validation system to acquire images under planned illumination for detailed validation.

4.2. Real-World Robotic Validation System

To validate the continuous planning policy in a physical environment, we constructed an automated acquisition system capable of unrestricted hemispherical sampling in Fig. 4. The system operates in a dark room to suppress ambient illumination and comprises (i) an industrial camera with a fixed viewpoint; (ii) a motorized dual-axis robotic arm (Arduino-controlled) that steers a point light source in arbitrary directions in the upper hemisphere; (iii) a specular calibration sphere positioned near the target object for light-direction calibration; and (iv) the target object. This setup provides the precise, repeatable directional control required by COIL-PS, enabling the capture of high-quality image sequences under planned illumination for quantitative assessment.

Imaging and Illumination Subsystems. The illumination is provided by a high-power LED mounted on a dual-axis robotic arm, which offers independent control over pitch and azimuth for precise angular positioning. A convex lens is placed in front of the LED to collimate the beam, enhancing directionality and irradiance uniformity to approximate a distant point source. The two rotation axes are designed to be concentric with the target object, ensuring that the LED moves on a spherical trajectory at a nearly constant radius of approximately 50 cm from the object. This concentric design maintains consistent irradiance across different poses and minimizes intensity variations unrelated to surface reflectance.

High Dynamic Range Imaging. We employ a Daheng MER-503-36U3C camera with a 50 mm lens, capturing 12-bit linear RAW images at 2448×2048 resolution; images are centrally cropped to mitigate vignetting. To preserve shadows and specular highlights, we acquire an exposure bracket (1–10 ms) for each illumination and merge it into a single HDR image [36].

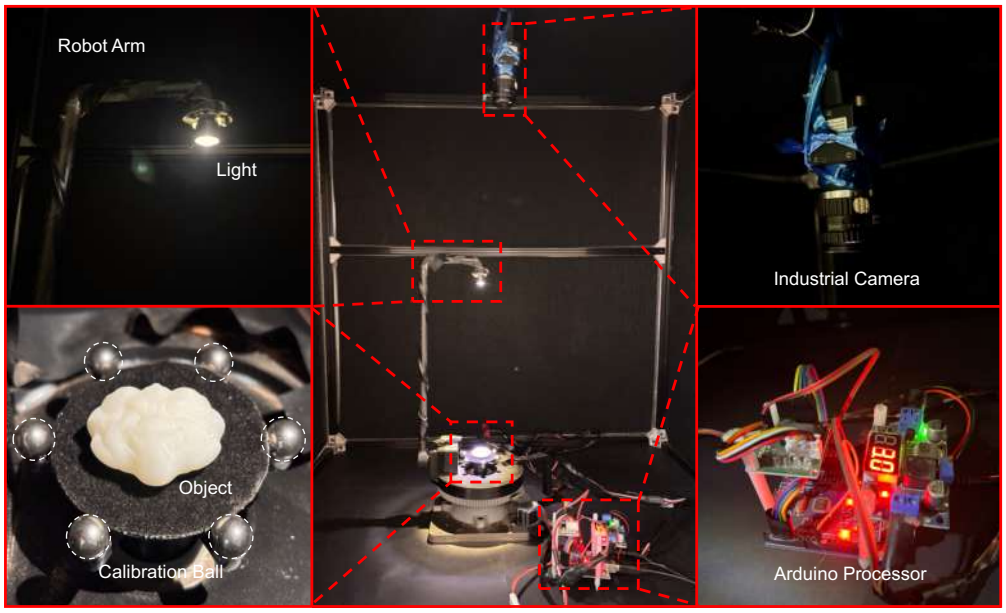


Fig. 4. Real-world robotic validation system. **(Center)**: The experimental setup features an Arduino-controlled dual-axis robotic arm capable of unrestricted hemispherical positioning of the light source around the target object, a fixed industrial camera, the target object on a platform, and control electronics. **(Top left)**: The collimated high-power LED mounted on the dual-axis arm approximates a point source. **(Bottom left)**: The target object ("Monster") and calibration spheres mounted on the stage. **(Top right)**: An industrial camera captures 12-bit RAW data for HDR processing. **(Bottom right)**: Arduino processor that drives arm motion and synchronizes image acquisition. This setup allows for the physical realization of the continuous light coordinates predicted by COIL-PS.

Ground-Truth Normal Acquisition. For ground-truth normal map acquisition, we use the calibrated intrinsic parameters together with a refined extrinsic pose to render the CAD mesh into a surface normal map in Blender [31] at the same resolution as the captured images. Although the CAD model is used to fabricate the 3D-printed object, small discrepancies may still arise due to printing tolerances, camera placement, and the physical scene setup, making perfect geometric alignment difficult to guaranty. Although the alignment is verified at sub-pixel precision, slight pose discrepancies remain unavoidable; therefore, the rendered normals should be regarded as approximate rather than absolute "ground truth".

Data Acquisition Pipeline. We use the "Monster" and the "Sphere" as the target objects. For the other IP methods, we capture 100 images in advance under uniform light sampling, which is comprehensive but time-consuming. In contrast, COIL-PS operates with a maximum of $T = 10$ planning steps: starting from an initial image, it sequentially predicts and acquires the next light direction. With its continuous formulation, COIL-PS commands the robotic arm to arbitrary positions on the upper hemisphere, enabling fine-grained and adaptive selection beyond predefined discrete lights. For each selected light direction, the arm is positioned, an exposure bracket is captured to form an HDR image, the observation is appended to the stack, and COIL-PS determines the next light direction until the maximum T steps are reached. The resulting set of light directions and HDR images is then used for error calculation (Eq. (2)) and normal estimation (Eq. (3)).

Table 1. Performance (MAE in degrees) for a fixed budget of $T = 10$ lights of COIL-PS versus other IP methods on synthetic dataset and semi-real benchmarks.

Dataset	Random	DC05 [16]	TK22 [17]	ReLeaPS [18]	LIPIDS [20]	LAC-PS [19]	COIL-PS (Ours)
Sculpture [21]	19.93	20.91	19.80	18.09	19.43	19.18	**17.15**
DiLiGenT [15]	14.62	14.20	14.12	13.98	14.06	14.20	**13.83**
DiLiGenT10^2 [35]	21.11	20.16	20.68	20.83	20.39	21.57	**17.99**

Table 2. Performance (MAE in degrees) for a fixed budget of $T = 10$ lights of COIL-PS versus other IP methods on the DiLiGenT [15] benchmark.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	**Average**
Random	3.30	9.24	13.83	8.13	26.25	16.48	28.84	8.46	14.63	17.03	14.62
DC05 [16]	4.49	9.02	13.92	8.27	**23.46**	15.80	27.96	8.17	13.96	16.97	14.20
TK22 [17]	3.61	8.99	13.40	8.16	24.40	16.24	28.07	7.98	13.79	16.55	14.12
ReLeaPS [18]	**3.14**	**8.55**	13.37	8.87	24.38	15.87	27.85	7.96	**13.54**	16.26	13.98
LIPIDS [20]	3.70	8.75	13.43	8.09	23.98	16.35	27.88	8.07	13.62	16.76	14.06
LAC-PS [19]	3.73	9.03	13.43	8.04	24.75	16.34	28.34	**7.87**	14.51	**15.94**	14.20
COIL-PS (Ours)	3.38	9.00	**12.92**	**7.75**	24.08	**15.70**	**27.12**	7.93	14.01	16.45	**13.83**

4.3. Results on Synthetic Dataset and Semi-Real Benchmarks

We evaluate the performance of our proposed method, COIL-PS, against previous state-of-the-art IP methods on both synthetic and semi-real benchmarks. This comparison validates our core premise: that an active and sequential continuous planning strategy can identify superior lighting configurations for a fixed budget. For synthetic evaluation, we employ the Sculpture dataset [21], while for real-world validation, we utilize two benchmarks from the DiLiGenT series: the standard DiLiGenT [15] and the more challenging DiLiGenT10^2 [35] benchmarks. All experiments are conducted under the strict budget of $T = 10$ lights. As summarized in Table 1, COIL-PS achieves the lowest average MAE across all datasets. It outperforms existing approaches, including random light selection, DC05 [16], TK22 [17], ReLeaPS [18], LIPIDS [20], and LAC-PS [19]. On the synthetic Sculpture dataset, our COIL-PS planning framework reduces the MAE to 17.15°, surpassing the next best method—ReLeaPS [18]—by 0.94°. This improvement directly demonstrates the advantage of continuous optimization over discrete candidate sampling. In semi-real evaluation, COIL-PS attains an MAE of 13.83° on DiLiGenT [15] and 17.99° on DiLiGenT10^2 [35]. The DiLiGenT [15] benchmark confirms its strong generalization to real-world objects with moderate shape-light interactions. Meanwhile, the challenging DiLiGenT10^2 [35] benchmark—with its more complex reflectances—highlights how the proposed COIL-PS method can robustly determine optimal light directions for real-world objects with different reflectance properties. These results demonstrate that our continuous illumination planning approach provides superior robustness and generalization under varying complexities by adaptively selecting the most informative lights with 10 planning steps.

To assess the performance of our proposed COIL-PS, we analyze each object in the DiLiGenT [15] benchmark, which contains 10 common objects featuring complex shapes and challenging reflectance. As shown in Table 2, COIL-PS achieves strong performance across all data in the DiLiGenT [15] benchmark, attaining the lowest average MAE of 13.83°. This result surpasses all previous illumination planning methods and attains the best results on several challenging objects, such as the "Buddha", "Cat", and "Goblet", demonstrating its robustness concerning these objects. As illustrated in Fig. 5, the visualization presents the qualitative results for the "Buddha" and "Cat" objects from the DiLiGenT [15] benchmark for different illumination planning methods under varying light budgets ($T = 3, 5, 10$). Across varying light budgets, COIL-PS exhibits markedly greater robustness than previous IP methods, particularly in deeply

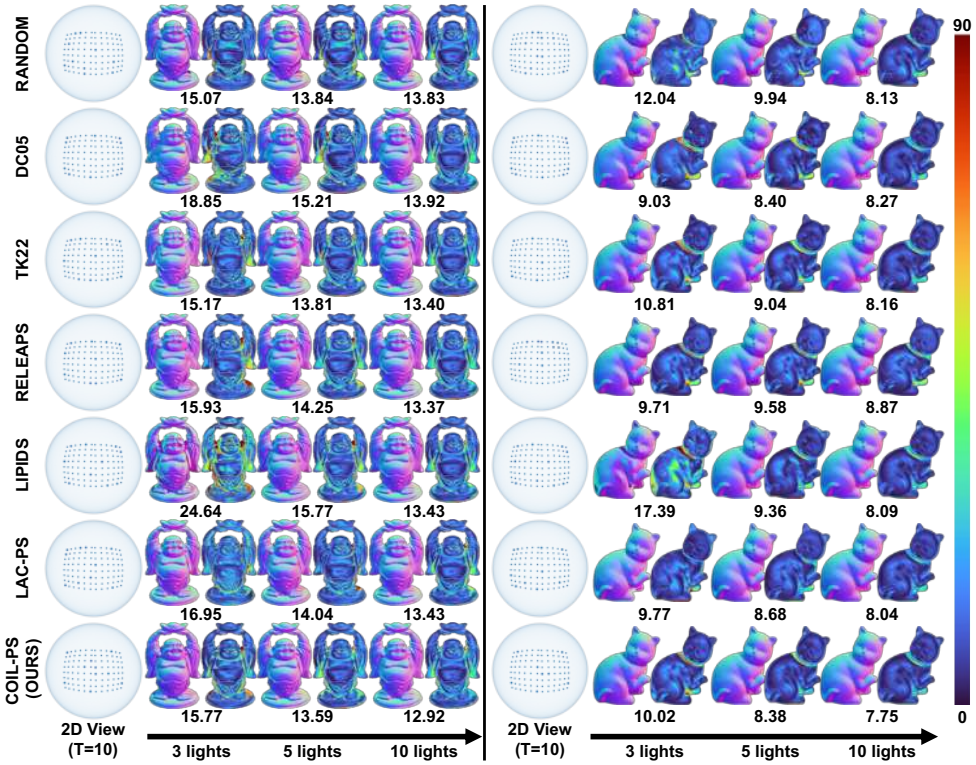


Fig. 5. Comparison of reconstruction results and corresponding error maps for the "Buddha" and "Cat" objects from the DiLiGenT [15] benchmark under a budget of $T = 10$ lights. The error maps demonstrate how our continuous IP (COIL-PS) better resolves fine-grained geometry and reduces errors in high-frequency regions compared to state-of-the-art discrete IP methods.

Table 3. Performance (MAE in degrees) of COIL-PS versus other IP methods on the "Monster" and "Sphere" objects captured using our custom-built real-world valudation system in $T = 10$ lights.

Dataset	Random	DC05 [16]	TK22 [17]	ReLeaPS [18]	LIPIDS [20]	LAC-PS [19]	COIL-PS (OURS)
Monster	9.34	9.04	8.51	9.06	9.05	9.58	**7.99**
Sphere	12.42	11.59	14.04	7.53	9.47	14.43	**6.87**

shadowed and geometrically complex regions. On the "Buddha", where strong inter-reflections and concave structures challenge the photometric cues, competing approaches (*e.g.*, TK22 [17], ReLeaPS [18], LAC-PS [19]) show limited improvement as the number of lights increases, whereas COIL-PS achieves reductions in error (15.77°→13.59°→12.92°) and reconstructs the deepest cavities, such as the folded robe and the base. On the "Cat", characterized by broad specular highlights and smooth curvature, several baselines remain unstable under sparse illumination, while COIL-PS improves steadily (10.02°→8.38°→7.75°). The 3-light and, in some cases, the 5-light settings yield suboptimal performance compared to other IP methods. This is attributable to the DiLiGenT [15] benchmark's lighting configuration, where lights are concentrated in a narrow central region, thereby constraining continuous illumination planning. To conclusively validate the advantages of COIL-PS, we performed a real-world evaluation using a custom-built system that provides full hemispherical directional illumination.

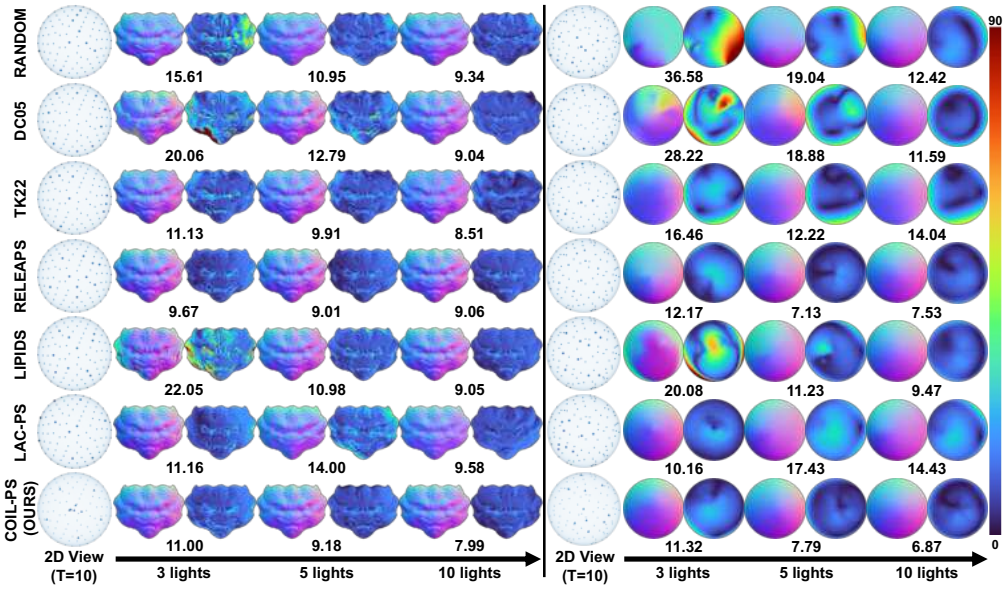


Fig. 6. Comparison of reconstruction results and corresponding error maps for the "Monster" and "Sphere" objects from the images captured under our real-world robotic validation system. This robotic validation system enables the precise, continuous illumination control required by our COIL-PS framework. By operating over a continuous domain—rather than a discrete grid as in other IP methods—our approach achieves better reconstruction performance on objects with complex reflectance.

4.4. Results on Real-World Data using Our Robotic Validation System

We evaluate our proposed COIL-PS method in real-world scenarios using a custom-built robotic validation system. The evaluation employs objects ("Monster" and "Sphere") that exhibit complex geometry within a lighting budget of $T = 10$ lights. As summarized in Table 3, COIL-PS achieves the lowest average MAE on both objects, outperforming all other IP methods. On the "Monster", our method attains an MAE of 7.99°, surpassing the next best result (TK22 [17] at 8.51°) by 0.52°. The improvement is more pronounced on the "Sphere" object, where COIL-PS reduces the error to 6.87°, outperforming the second best IP method (ReLeaPS [18] at 7.53°) by 0.66°. Notably, methods such as LAC-PS [19] and TK22 [17] exhibit high variance, performing well on one object but poorly on another, while COIL-PS demonstrates consistent robustness. These results underscore the effectiveness of our continuous illumination planning strategy in identifying optimal lighting configurations that generalize across diverse object properties, leading to state-of-the-art reconstruction accuracy.

The qualitative results in Fig. 6 demonstrate the robustness of our method, COIL-PS, across two real objects with diverse geometric complexities. Our proposed COIL-PS produces more stable and accurate surface normal predictions than other IP methods under varying lighting conditions (3, 5, and 10 lights). On the geometrically complex "Monster" object, COIL-PS excels in recovering deep concave regions—most notably the shadowed eye sockets and interior facial cavities. With only 3 lights, COIL-PS achieves a competitive angular error of 11.00°, slightly worse than ReLeaPS [18]. Performance improves progressively with additional lights (9.18° at 5 lights, 7.99° at 10 lights), confirming its ability to leverage light cues on continuous coordinates for fine-scale geometry reconstruction. On the 'Sphere" object, COIL-PS again shows notable stability (11.32° under 3 lights), whereas other methods suffer severe orientation drift (*e.g.*, random light selection: 36.58°, DC05 [16]: 28.22°, TK22 [17]: 16.46°, and LIPIDS [20]: 20.08°). As the number of lights increases, COIL-PS converges smoothly toward high-fidelity

Table 4. Impact of network components on reconstruction accuracy (MAE in degrees) with $T = 10$ lights on Sculpture [21], DiLiGenT [15], and DiLiGenT10^2 [35] using the least-square estimation [5]. Removing the Outlier-Aware Gating (w/o Gating) or the Transformer (w/o Transformer) leads to consistent performance degradation, validating the necessity of the proposed architecture.

Dataset	w/o Gating	w/o Transformer	Ours
Sculpture [21]	17.73	17.79	**17.01**
DiLiGenT [15]	13.96	14.04	**13.83**
DiLiGenT10^2 [35]	18.72	19.06	**17.99**

normal maps (7.79° at 5 lights, 6.87° at 10 lights), effectively suppressing the artifacts in ReLeaPS [18], TK22 [17], and LAC-PS [19]. Although the 3-light and 5-light setups inherently limit accuracy due to sparse lighting, COIL-PS still produces high-quality reconstructions, recovering subtle shadowed details that other IP methods miss and maintaining reliability even when extended to 10 lights.

4.5. Ablation Studies

We perform an ablation study to evaluate the contributions of two key components in COIL-PS. The first variant, "COIL-PS w/o Gating", removes the outlier-aware gating mechanism, which mitigates photometric outliers such as shadows and specularities. The second variant, "COIL-PS w/o Transformer", omits the transformer encoder and its positional encoding, thereby disabling its capacity to model temporal dependencies in the illumination sequence. The ablation studies in Table 4 confirm the necessity of each component in COIL-PS, as the removal of any single part degrades the performance of our proposed COIL-PS, thereby validating our architectural design.

5. Conclusion

In this work, we challenged the widely adopted assumption of discrete-space IP in PS. We introduced COIL-PS, the first framework to formulate light selection as a continuous regression problem. Our system leverages a feedback-driven policy that steers illumination with arbitrary angular precision. Across synthetic data and the semi-real benchmarks, including DiLiGenT [15] and DiLiGenT10^2 [35], COIL-PS achieves state-of-the-art or competitive accuracy with a budget of 10 lights. The validation of our proposed COIL-PS on a custom-built real-world system also proves that our continuous sampling strategy yields superior metrological performance. We attribute these performance gains to two key factors: first, operating directly in continuous light space, which eliminates quantization artifacts and allows for fine-grained directional refinement; and second, employing a closed-loop illumination strategy conditioned on the evolving reconstruction, which enables adaptive correction and improves the handling of non-Lambertian effects. Although our implementation assumes distant lighting and near-orthographic imaging, COIL-PS establishes a principled framework for practical PS by coupling continuous light-space reasoning with reconstruction-driven acquisition. Promising extensions include relaxing these assumptions, modeling complex BRDFs, and jointly illumination planning for dynamic scenarios.

References

1. S. Han, J. Park, and D. Cho, "Minifying photometric stereo via knowledge distillation-based feature translation," Opt. Express **30**, 38284–38297 (2022).
2. D. Yan, P. Zhang, R. Song, *et al.*, "BPSCR: 3d contour-reconstruction method for sheet metal based on binocular photometric stereo vision," Opt. Express **33**, 25980–25992 (2025).
3. Y. Braun and H. Guterman, "Light invariant photometric stereo," Opt. Express **31**, 5200–5214 (2023).
4. K. Zhang, X. Zhao, Y. Wen, and D. Li, "PSLFM: a single-frame uncalibrated photometric stereoscopic light field measurement scheme based on dense convolutional neural networks," Opt. Express **33**, 3082–3100 (2025).

5. R. J. Woodham, "Photometric method for determining surface orientation from multiple images," Opt. Eng. **19**, 139–144 (1980).
6. Y. Mukaigawa, Y. Ishii, and T. Shakunaga, "Analysis of photometric factors based on photometric linearization," J. Opt. Soc. Am. A, Opt. Image Sci. Vis. **24**, 3326–34 (2007).
7. Y. M. Satoshi Ikehata, David Wipf and K. Aizawa, "Robust photometric stereo using sparse regression," in *IEEE Computer Vision and Pattern Recognition,* (2012).
8. K.-L. Tang, C.-K. Tang, and T.-T. Wong, "Dense photometric stereo using tensorial belief propagation," in *IEEE Computer Vision and Pattern Recognition,* (2005).
9. H. Santo, M. Samejima, Y. Sugano, *et al.*, "Deep photometric stereo network," in *IEEE International Conference on Computer Vision Workshops,* (2017).
10. S. Ikehata, "CNN-PS: CNN-based photometric stereo for general non-convex surfaces," in *European Conference on Computer Vision,* (2018).
11. Y. Ju, K.-M. Lam, Y. Chen, *et al.*, "Pay attention to devils: A photometric stereo network for better details," in *International Joint Conference on Artificial Intelligence,* (2021).
12. G. Chen, K. Han, and K.-Y. K. Wong, "PS-FCN: A flexible learning framework for photometric stereo," in *European Conference on Computer Vision,* (2018).
13. G. Chen, K. Han, B. Shi, *et al.*, "SDPS-Net: Self-calibrating deep photometric stereo networks," in *IEEE Computer Vision and Pattern Recognition,* (2019).
14. T. Taniai and T. Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *International Conference on Machine Learning,* (2018).
15. B. Shi, Z. Wu, Z. Mo, *et al.*, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," in *IEEE Computer Vision and Pattern Recognition,* (2016).
16. O. Drbohlav and M. Chantler, "On optimal light configurations in photometric stereo," in *IEEE International Conference on Computer Vision,* (2005).
17. H. Tanikawa, R. Kawahara, and T. Okabe, "Online illumination planning for shadow-robust photometric stereo," in *International Workshop on Frontiers of Computer Vision,* (2022).
18. J. H. Chan, B. Yu, H. Guo, *et al.*, "ReLeaPS: Reinforcement learning-based illumination planning for generalized photometric stereo," in *IEEE International Conference on Computer Vision,* (2023).
19. W. Meng, H. Han, X. Lu, *et al.*, "LAC-PS: A light direction selection policy under the accuracy constraint for photometric stereo," IEEE Trans. on Circuits Syst. for Video Technol. (2025).
20. A. Tiwari, M. Sutariya, and S. Raman, "LIPIDS: Learning-based illumination planning in discretized (light) space for photometric stereo," in *IEEE Winter Conference on Applications of Computer Vision,* (2025).
21. O. Wiles and A. Zisserman, "SILNet: Single-and multi-view reconstruction by learning from silhouettes," arXiv preprint arXiv:1711.07888 (2017).
22. A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems,* (2017).
23. C. Yu, Y. Seo, and S. W. Lee, "Photometric stereo from maximum feasible lambertian reflections," in *European Conference on Computer Vision,* (2010).
24. D. Miyazaki, K. Hara, and K. Ikeuchi, "Median photometric stereo as applied to the segonko tumulus and museum objects," Int. J. Comput. Vis. **86**, 229–242 (2009).
25. Y. M. Satoshi Ikehata, David P. Wipf and K. Aizawa, "Photometric stereo using sparse bayesian regression for general diffuse surfaces," IEEE Trans. on Pattern Anal. Mach. Intell. **36**, 1078–1091 (2014).
26. L. Wu, A. Ganesh, B. Shi, *et al.*, "Robust photometric stereo via low-rank matrix completion and recovery," in *Asian Conference on Computer Vision,* (2010).
27. F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla, "PX-Net: Simple and efficient pixel-wise training of photometric stereo networks," in *IEEE International Conference on Computer Vision,* (2021).
28. Á. González, "Measurement of areas on a sphere using fibonacci and latitude–longitude lattices," Math. Geosci. **42**, 49–64 (2010).
29. Unity Technologies, "Unity Game Engine," (2025). Available at https://unity.com.
30. Epic Games, "Unreal Engine," (2025). Available at https://www.unrealengine.com/.
31. Blender Online Community, "Blender: A 3D modelling and rendering package," (2025). Version 5.0, Available at https://www.blender.org.
32. M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob, "Mitsuba 2: A retargetable forward and inverse renderer," ACM Trans. on Graph. (ToG) **38**, 1–17 (2019).
33. B. Yu, S. Yang, X. Cui, *et al.*, "MILO: Multi-bounce inverse rendering for indoor scene with light-emitting objects," IEEE Trans. on Pattern Anal. Mach. Intell. **45**, 10129–10142 (2023).
34. B. Shi, Z. Mo, Z. Wu, *et al.*, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," IEEE Trans. on Pattern Anal. Mach. Intell. **41**, 271–284 (2019).
35. J. Ren, F. Wang, J. Zhang, *et al.*, "DiLiGenT10^2: A photometric stereo benchmark dataset with controlled shape and material variation," in *IEEE Computer Vision and Pattern Recognition,* (2022).
36. P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," Seminal Graph. Pap. Push. Boundaries **2**, 643–652 (2023).