

Geometry Meets Light: Leveraging Geometric Priors for Universal Photometric Stereo Under Limited Multi-Illumination Cues

King-Man Tam¹, Satoshi Ikehata^{2,3}, Yuta Asano², Zhaoyi An¹, Rei Kawakami¹

¹Institute of Science Tokyo

²National Institute of Informatics

³Denso IT Laboratory

tam.k.4f46@m.isct.ac.jp, sikehata@nii.ac.jp

Abstract

Universal Photometric Stereo is a promising approach for recovering surface normals without strict lighting assumptions. However, it struggles when multi-illumination cues are unreliable, such as under biased lighting or in shadows or self-occluded regions of complex in-the-wild scenes. We propose GeoUniPS, a universal photometric stereo network that integrates synthetic supervision with high-level geometric priors from large-scale 3D reconstruction models pretrained on massive in-the-wild data. Our key insight is that these 3D reconstruction models serve as visual-geometry foundation models, inherently encoding rich geometric knowledge of real scenes. To leverage this, we design a Light-Geometry Dual-Branch Encoder that extracts both multi-illumination cues and geometric priors from the frozen 3D reconstruction model. We also address the limitations of the conventional orthographic projection assumption by introducing the PS-Perp dataset with realistic perspective projection to enable learning of spatially varying view directions. Extensive experiments demonstrate that GeoUniPS delivers state-of-the-arts performance across multiple datasets, both quantitatively and qualitatively, especially in the complex in-the-wild scenes.

Code — <https://github.com/marcotam2002/geounips>

Introduction

Photometric Stereo (PS) (Woodham 1980) is a method for recovering high-fidelity surface normals from multiple images captured under varying illumination with a fixed camera. Historically, the development of PS can be seen as a gradual relaxation of assumptions about the lighting conditions. Traditional PS methods relied on physically based inverse rendering with calibrated directional lighting and specific BRDFs (e.g., Lambertian) (Ikehata et al. 2012; Shi et al. 2014). Early learning-based methods showed that normals of non-Lambertian, non-convex surfaces could be directly regressed from images in an uncalibrated setup, but they still assumed directional lighting models (Chen et al. 2020; Sarno et al. 2022). More recently, physics-free universal PS methods (Ikehata 2022, 2023; Ikehata and Asano 2024) have removed the need for explicit lighting models, enabling use with arbitrary uncalibrated light sources.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

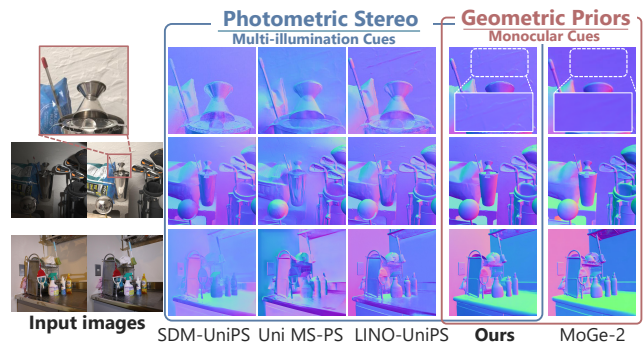


Figure 1: Our method effectively leverages geometric priors from pretrained 3D reconstruction model, achieving more plausible normal map recovery in challenging scenes with complex backgrounds and limited lighting variation. Compared to SoTA monocular normal prediction models (e.g., MoGe-2 (Wang et al. 2025c)), our approach captures finer surface details by incorporating multi-illumination cues.

While recent advances have eliminated many assumptions about lighting, one critical premise remains: *each surface point is assumed to be observed under sufficiently diverse, well-distributed lighting*. However, this condition often fails in real-world settings. Due to the practical difficulty of controlling light sources, some regions may receive rich and varied illumination, while others receive far less. As illustrated in Fig. 1 (bottom), the scene was illuminated using a moving handheld flashlight from the front. Performance in poorly illuminated areas degrades significantly, as also discussed in (Ikehata 2023); the degradation is even more pronounced in regions with complex geometry (e.g., containers), material (e.g., mirror surface) and textures (e.g., bottles), where limited lighting cues make normal estimation more difficult. This limitation fundamentally arises because photometric stereo relies on multi-illumination cues induced by changes in illumination as its primary cue; when these variations are unreliable, the method lacks a mechanism to compensate.

A straightforward way to handle this can be to train models on large-scale in-the-wild multi-illumination datasets to learn monocular geometric priors, following recent 3D reconstruction studies. For example, recent feedforward mul-

tiview 3D reconstruction models trained on millions of in-the-wild images, e.g., (Wang et al. 2024, 2025a), have shown that even with a single input image, a network can recover a plausible 3D shape, despite being trained with multiple images, indicating that these models have learned high-level monocular priors beyond low-level multiview photometric constraints.

However, applying this strategy to photometric stereo is extremely challenging. Acquiring ground-truth normal maps for real scenes with the resolution and fidelity required by PS is prohibitively expensive, and it is even more difficult to cover the vast combinations of lighting conditions, surface properties and camera settings. Consequently, existing PS models are exclusively trained on clean synthetic datasets, whose statistics differ markedly from those of real scenes which naturally encode rich geometric and contextual priors (e.g., building façades tend to be piecewise planar). Without exposure to such real-world regularities, PS models could largely rely on shading variations, leaving poorly illuminated regions without meaningful guidance. This raises a crucial question: *How can we acquire high-level geometric priors in a synthetic photometric stereo training pipeline?*

In this paper, we present *GeoUniPS*, a geometry-aware universal photometric stereo network that combines synthetic supervision with high-level priors from large-scale 3D reconstruction models (e.g., VGGT (Wang et al. 2025a)). Our key insight is that 3D reconstruction models, pretrained on massive in-the-wild datasets, act as visual-geometry foundation models and inherently encode rich geometric knowledge of real scenes. We find that by injecting their features into the photometric stereo pipeline, models can leverage geometric priors unattainable through purely synthetic multi-illumination training. Even with limited illumination variation, these priors regularize surface normal estimation, yielding more reliable and globally consistent results as illustrated in Fig. 1 (top).

Technically, we propose a novel dual-branch encoder: one branch extracts lighting-aware features through synthetic supervision, while the other captures lighting-invariant, high-level geometric features from a frozen 3D reconstruction model. These complementary cues are fused into a unified representation, which the decoder leverages to produce context-aware, geometry-faithful normals. Unlike prior methods that rely solely on multi-illumination cues, *GeoUniPS* benefits from the pretrained model’s embedded geometric priors, enabling contextually valid estimations even when multi-illumination cues are limited.

We further address a gap in existing training datasets. Previous photometric stereo datasets typically assume orthographic projection, whereas real world setups operate under perspective projection. To bridge this gap, we construct *PS-Perp*, the first synthetic training dataset with realistic perspective projection, enabling the network to learn spatially varying view directions.

In our evaluation, we demonstrate that *GeoUniPS* enriched with priors from 3D reconstruction models, achieves more plausible results across various datasets. These include standard photometric stereo benchmarks under single directional lighting (Shi et al. 2016; Roberto Mecca

and Cipolla 2021), as well as real world multi-illumination datasets (Murrman et al. 2019).

Related Work

Photometric Stereo: Photometric stereo has a long history in computer vision, estimating surface normals of static scenes from images under varying illumination. The classic method (Woodham 1980) assumes Lambertian, convex surfaces, and known directional lighting in a darkroom. Later works extended it to handle non-Lambertian scenes, using robust techniques that treat non-Lambertian effects as outliers (Wu et al. 2010; Ikehata et al. 2012) or explicitly use non-Lambertian BRDF (Alldrin and Kriegman 2007; Goldman et al. 2010).

With the progress of deep learning, data-driven photometric stereo emerged under calibrated lighting, targeting non-Lambertian, non-convex surfaces via observation maps (Ikehata 2018), set-pooling (Chen, Han, and Wong 2018), graph neural networks (Yao et al. 2020), and Transformers (Ikehata 2021). Meanwhile, neural inverse rendering methods (Taniai and Maehara 2018; Li and Li 2022) adopted physics-guided, unsupervised learning to estimate normals without supervision. Uncalibrated settings were also addressed by self-calibrating networks predicting light and normals sequentially (Chen et al. 2019, 2020; Kaya et al. 2021).

However, these methods assume simplified lighting models, limiting their applicability under complex real-world lighting conditions. To address this, universal photometric stereo (Ikehata 2022) has been proposed, aiming to learn lighting representations directly from images without restrictive lighting assumptions. *SDM-UniPS* (Ikehata 2023) eliminated the need for masks and enabled high-resolution normal recovery using a pixel-sampling Transformer. *UniMS-PS* (Hardy, Quéau, and Tschumperlé 2024) extended this approach to a multi-scale architecture, while *SpectraMPS* (Ikehata and Asano 2024) introduced the first universal multispectral photometric stereo networks for dynamic surfaces. Most recently, *LINO-UniPS* (Li et al. 2025) decoupled lighting from geometry, enhancing detail preservation through wavelet-based processing and a gradient-aware loss.

Despite recent progress, photometric stereo remains challenged by limited or biased lighting, where multi-illumination cues are not reliable. Its reliance on synthetic training data, which lacks the geometric context of real-world scenes, further limits generalization. These issues highlight the need for techniques to induce stronger, high-level geometric priors.

Feedforward 3D Reconstruction Models: Recent feed-forward 3D reconstruction models have shifted multi-view reconstruction from traditional SfM + MVS pipelines to end-to-end neural inference. *DUSt3R* (Wang et al. 2024) demonstrates that a Transformer-based model can reconstruct dense point maps from unposed image pairs in a single pass. Building on this, *MASt3R* (Leroy, Cabon, and Revaud 2024) improves dense correspondence quality in wide-baseline scenarios. Multi-view extensions such as *VGGT* (Wang et al. 2025a) and *MV-DUSt3R+* (Tang et al. 2025) further scale this paradigm. For instance, *VGGT* can

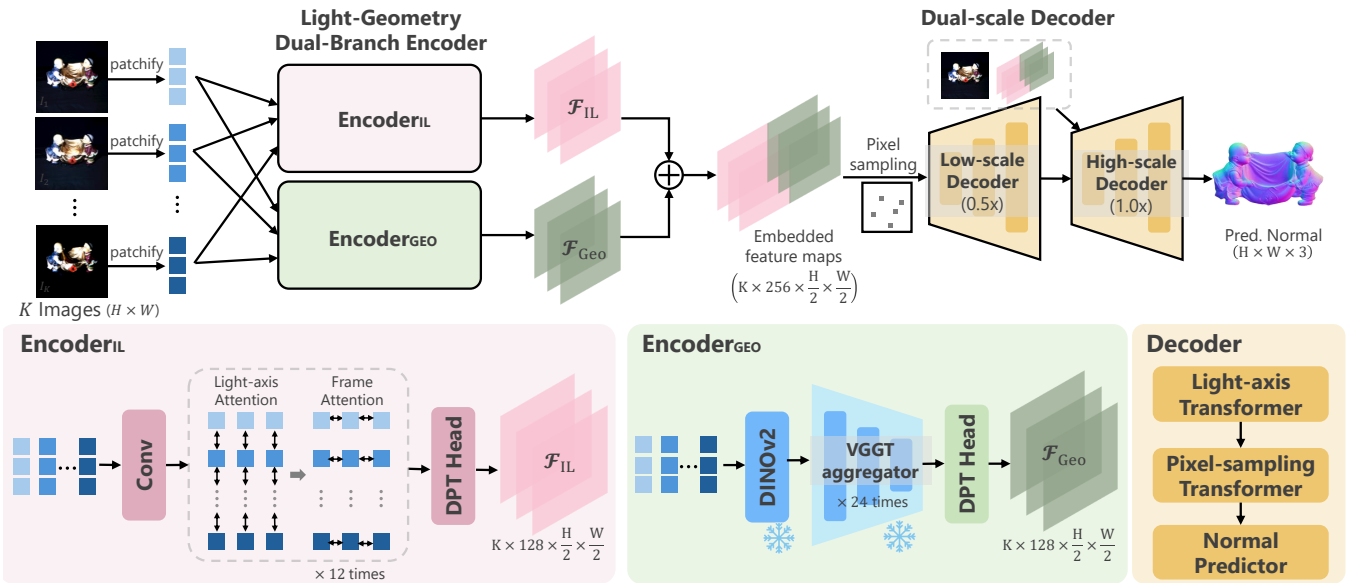


Figure 2: Overview of our GeoUniPS architecture. Given multiple input images captured under different lighting conditions, the Light-Geometry Dual-Branch Encoder extracts both light-variant features from multi-illumination cues ($\text{Encoder}_{\text{IL}}$) and geometric features from the pretrained VGGT aggregator ($\text{Encoder}_{\text{Geo}}$). These features are concatenated with the input images using an MLP-based embedding, after which the Dual-Scale Normal Decoder performs pixel-wise normal regression at sampled locations.

process anywhere from a single image to hundreds of views in under a second, predicting camera intrinsics/extrinsics, depth maps, point maps, and even point tracks in a feed-forward manner. Crucially, VGGT shows that even a single image can suffice to reconstruct dense geometry, confirming that these networks encode high-level geometric knowledge beyond simple matching cues. Trained on large-scale in-the-wild datasets such as MegaDepth, CO3D-v2, ScanNet, and DL3DV, these models generalize robustly and demonstrate zero-shot performance in novel and challenging scenarios. This provides strong evidence that feed-forward 3D reconstruction models have learned rich priors about scene geometry directly from large-scale data.

Method

We propose **GeoUniPS**, the first universal photometric stereo network that integrates geometric priors from pretrained 3D reconstruction models with multi-illumination cues. We begin by formulating the universal photometric stereo problem and algorithm, then describe our architecture and training dataset design. We address two main challenges in real-world scenario: (1) limited illumination cues, such as weak illumination variations, shadows, or self-occlusion, which degrade surface normal estimation; and (2) the orthographic projection assumption, which fails to adapt to perspective scenes in real-world settings. To overcome them, we propose both architectural and data-driven solutions.

On the architectural side, we introduce a *Light-Geometry Dual-Branch Encoder* that extracts both high-level geometric priors and multi-illumination cues. Unlike previous methods that rely solely on multi-illumination cues, we in-

corporate geometric guidance from a pretrained foundation model (e.g., VGGT (Wang et al. 2025a)) to compensate for limited multi-illumination cues.

To adapt our model to general perspective scenes, we first demonstrate the universal photometric stereo networks trained on a large scale synthetic dataset rendered under *perspective projection*, comprising over 60,000 scenes with diverse shapes, materials, and focal lengths. This dataset exposes the model to realistic perspective distortions.

Problem Statement and Algorithm Formulation

Universal photometric stereo (Ikehata 2022) is a variant of the photometric stereo task where no information about the lighting conditions is available (i.e., no calibration nor prior lighting models), unlike conventional calibrated (Woodham 1980) or uncalibrated (Chen et al. 2019) PS tasks. Specifically, given a set of K images $\mathcal{I} = \{I_k\}_{k=1}^K$,

$$I_k \in \mathbb{R}^{H \times W \times 3}$$

, captured by a fixed camera under arbitrary lighting conditions (commonly referred to as multi-illumination images), the goal of universal photometric stereo is to recover a pixel-accurate surface normal map $N \in \mathbb{R}^{H \times W \times 3}$. Importantly, the number of images K is arbitrary at test time. An optional object mask $M \in \mathbb{R}^{H \times W}$ is often provided, but our method does not assume the availability of this mask at inference (yet used during training), as it introduces additional cost.

Most existing universal photometric stereo networks (Ikehata 2022, 2023; Ikehata and Asano 2024; Li et al. 2025) are characterized by a two-stage design: an encoder that extracts image-level feature maps from individual images with intra-

and inter-frame interactions, and a decoder that recovers the surface normal at randomly sampled pixel locations using the extracted features and the original image information. Following them, GeoUniPS also has the same two-stage design.

The role of the encoder is to extract K feature maps from \mathcal{I} . Formally,

$$\mathcal{F} = \{F_k\}_{k=1}^K = \text{Encoder}(\mathcal{I}), \quad F_k \in \mathbb{R}^{H' \times W' \times C}, \quad (1)$$

where C is the feature dimension, and $H' \times W'$ is the feature map size, typically smaller than $H \times W$.

Given P spatial coordinates $\mathcal{X} = \{\mathbf{x}_p\}_{p=1}^P$ where $\mathbf{x}_p \in [1, H] \times [1, W]$, the decoder predicts the surface normal $\mathbf{n}_p \in \mathbb{R}^3$ at each position \mathbf{x}_p from feature maps and images. Formally,

$$\{\mathbf{n}_p\}_{p=1}^P = \text{Decoder}(\mathcal{F}(\mathcal{X}\downarrow), \mathcal{I}(\mathcal{X})), \quad (2)$$

where $\mathcal{X}\downarrow$ represents the downsampled coordinates corresponding to the resolution change from $H \times W$ to $H' \times W'$. The encoder and decoder are trained solely with normal vector supervision on a large-scale synthetic dataset.

This two-stage design is inspired by traditional photometric stereo, where each image is paired with lighting information (e.g., light direction vectors), and surface normals are estimated by aggregating these pairs per pixel. In universal photometric stereo networks, this role is replaced by features extracted from images. The encoder’s feature maps thus serve as substitutes for lighting; SDM-UniPS (Ikehata 2023) refers to them as the Global Lighting Context (GLC).

Predicting normals at sparse pixels reduces memory and aligns with the core idea of estimating per-pixel normals from illumination variation without spatial context. Some variants incorporate spatial cues, such as patch-level features (Ikehata and Asano 2024) or coarse-to-fine strategies (Hardy, Quéau, and Tschumperlé 2024).

To isolate the contribution of geometric priors from pretrained models, *we follow this simple and established framework, minimizing confounding factors and ensuring that observed improvements are attributable to the geometric priors themselves.* The following sections detail the encoder and decoder design in GeoUniPS.

Light-Geometry Dual-Branch Encoder

We design our encoder to capture both illumination-aware and illumination-invariant priors. Specifically, the encoder consists of two branches:

- **Encoder_{Geo}**: This branch leverages high-level geometric priors that are invariant to illumination. We use the encoder from a pretrained 3D reconstruction model trained on large-scale in-the-wild data, followed by a learnable projector that adapts its output to our task. The pretrained encoder is frozen during training to preserve its geometric knowledge, which is expected to complement the illumination priors.
- **Encoder_{IL}**: This branch captures a multi-illumination prior by embedding shading variations across images taken under different lighting conditions from a fixed

viewpoint. It follows the motivation of conventional universal photometric stereo networks. This information is essential for the decoder to recover fine-grained surface normals from multi-illumination images.

The combined feature representation is obtained by simply concatenating the outputs of the two branches:

$$\mathcal{F} = \text{Concat}(\text{Encoder}_{\text{Geo}}(\mathcal{I}), \text{Encoder}_{\text{IL}}(\mathcal{I})), \quad (3)$$

where $\text{Concat}(\cdot)$ denotes channel-wise concatenation.

Encoder_{Geo} consists of a frozen encoder of a pretrained 3D reconstruction model, followed by a learnable projector. We adopt VGGT (Wang et al. 2025a) for its strong zero-shot geometry reconstruction capabilities. The projector is implemented as a DPT-head (Ranftl, Bochkovskiy, and Koltun 2021). Although VGGT was originally designed for multi-view reconstruction, it has been shown to perform well even with a single image by leveraging strong monocular geometric priors.

GeoUniPS uses VGGT’s aggregator without its decoders. As shown in Fig. 2, each input image is resized or cropped to a resolution divisible by 14, in accordance with VGGT’s architectural constraints. After preprocessing (i.e., DINOv2 normalization) and tokenization via DINOv2, the image tokens are passed to VGGT’s aggregator, which comprises 24 layers of alternating frame and global attention. We extract tokens from layers [4, 11, 17, 23] and fuse them using a learnable DPT head, projecting the VGGT features into a 128-dim feature map at a $2\times$ downsampled resolution, denoted as $\mathcal{F}_{\text{Geo}} \in \mathbb{R}^{K \times \frac{H}{2} \times \frac{W}{2} \times 128}$.

Encoder_{IL} adopts a Transformer-based architecture similar to VGGT’s aggregator with a DPT head, but replaces the DINOv2 tokenizer with a lightweight convolution layer to better capture fine-grained local patterns. The number of layers is reduced, as this encoder learns priors limited to smaller synthetic datasets. All parameters are trained from scratch.

Inter-image interaction is key to extracting illumination-aware features in this branch. Following prior universal photometric stereo methods, we replace VGGT’s full attention with light-axis attention. Note that, removing inter-image attention from Encoder_{Geo} showed negligible effect, likely because VGGT’s aggregator extracts strong geometric priors even from multi-illumination images as the view is fixed.

As in Encoder_{Geo}, each input is resized or cropped. After max-val normalization (per SDM-UniPS (Ikehata 2023)), the convolutional tokenizer processes images into tokens. These are passed to a modified VGGT’s aggregator with 12 layers, alternating between frame attention (intra-image) and light-axis attention (inter-image at the same spatial position). This design enables joint reasoning over spatial context and illumination differences. We extract tokens from layers [2, 5, 8, 11] and fuse them via a DPT head into a 128-dim feature map at $2\times$ downsampled resolution, denoted as $\mathcal{F}_{\text{IL}} \in \mathbb{R}^{K \times \frac{H}{2} \times \frac{W}{2} \times 128}$.

Finally, the embedded feature maps are given as

$$\mathcal{F} = \text{Concat}(\mathcal{F}_{\text{Geo}}, \mathcal{F}_{\text{IL}}). \quad (4)$$

Pixel-Sampling-Based Normal Decoder

While our primary focus is on the encoder, which leverages pretrained 3D reconstruction models, an effective normal

decoder is also essential for recovering detailed surface normals. We explore several decoder variants, including different embeddings of \mathcal{I} (i.e., pixel (Ikehata 2023), patch (Ikehata and Asano 2024), and MLP-based (Li et al. 2025)) and architectural designs (single-scale (Ikehata 2023; Li et al. 2025) vs. dual-scale (Ikehata and Asano 2024)). Although these variants achieve comparable benchmark scores, we find that the dual-scale decoder with MLP-based embedding provides the best trade-off between geometric detail and efficiency. Therefore, we adopt this design for our decoder.

Decoder_{GeoUniPS}($\mathcal{F}(\mathcal{X}\downarrow), \mathcal{I}(\mathcal{X})$) predicts surface normals at randomly sampled pixel locations \mathcal{X} from extracted features \mathcal{F} and input images \mathcal{I} . Henceforth, we treat the features at \mathcal{X} as tokens and apply Transformer layers to them. As shown in SDM-UniPS, the number of sampled pixels directly influences normal estimation accuracy, with larger sample sizes generally yielding better results. To balance performance with training efficiency, we sample 2,048 pixels during training and increase this to 10,000 pixels during inference. This sampling size can be further raised when memory permits, potentially leading to improved accuracy.

First, low-scale normals are estimated from $\mathcal{F}(\mathcal{X}\downarrow)$ using five 256-dimensional light-axis Transformers, followed by a 384-dimensional light-axis Transformer with Pooling-by-Multihead-Attention (PMA) (Lee et al. 2019) to aggregate features along the light-axis. Two 384-dimensional pixel-sampling Transformers (Ikehata 2023) then apply self-attention across spatial locations to enhance spatial coherence. Finally, a two-layer MLP (384→192→3) predicts the low-frequency normals, which are normalized to unit length.

For high-scale refinement, $\mathcal{I}(\mathcal{X})$ is passed through a 256-dimensional MLP with two LayerNorms (3→256), embedding RGB values into a high-dimensional space. The embedded features are concatenated with $\mathcal{F}(\mathcal{X}\downarrow)$, processed by five 256-dimensional light-axis Transformers (512→256), and aggregated along the light-axis via a 384-dimensional PMA (256→384). These features are fused with the low-scale normals into 387-dimensional representations. Spatial coherence is further enhanced by two 384-dimensional pixel-sampling Transformers, followed by a final MLP (387→384→192→3) that predicts the normals, which are normalized to unit length. The final normal map is reconstructed by aggregating predictions over all locations.

Training Loss: The training loss was computed using the Mean Squared Error (MSE) loss function to measure the ℓ_2 error between the predicted and ground truth surface normal vectors. This loss was calculated at both scales and then summed.

PS-Perp: Perspective Synthetic Training Dataset

To improve generalization to real-world scenes captured by perspective cameras, we introduce PS-Perp, the first large-scale synthetic dataset for Universal Photometric Stereo constructed using a perspective projection model. While existing datasets such as PS-Wild (Ikehata 2022), PS-Mix (Ikehata 2023), LINO-UNIPS (Li et al. 2025), and others (Hardy, Quéau, and Tschumperlé 2024; Yamaguchi et al. 2025) have contributed variations in scene and material complexity, they are all rendered under the orthographic camera

assumption. This limits their ability to train models that generalize to scenes exhibiting strong perspective distortion.

Unlike prior work, PS-Perp is rendered with a perspective camera using Blender’s Cycles renderer. Focal lengths are sampled from a broad range (20–1000mm) to simulate varying levels of distortion: shorter focal lengths (e.g., 20–70mm) induce strong perspective effects, while longer ones (e.g., 70-1000mm) approximate weak perspectives. In total, the dataset consists of 60,297 multi-object scenes, with 44,220 scenes rendered using focal lengths below 70mm to emphasize perspective diversity, and the remaining using longer focal lengths to include weakly distorted cases. Each scene is rendered into 10 16-bit images at 512×512 resolution under randomized combinations of directional, point, and environment lighting, following the PS-Mix (Ikehata 2023) pipeline. By sharing the same asset library, scene composition strategy, and lighting setup as PS-Mix, our dataset ensures compatibility while significantly expanding the coverage of camera models. PS-Perp enables training on a continuum from strongly to weakly distorted perspective images, bridging the gap between synthetic training data and realistic test conditions.

Since perspective cameras with long focal lengths cannot fully replicate orthographic views, we train our model using a combination of PS-Perp and PS-Mix (Ikehata 2023) to cover both projection types. Some samples of training data are provided in the appendix.

Results

Implementation Details and Computational Time

As described, our network was trained from scratch on a combination of PS-Perp and PS-Mix (Ikehata 2023). The model is implemented in PyTorch and trained on 4 NVIDIA H100 GPUs over 6 days using the AdamW (Kingma and Ba 2014) optimizer, with an initial learning rate of 1e-4, a weight decay of 0.05, and step decay (multiplied by 0.8 every 10 epochs). A linear warm-up is applied during the first epoch. For numerical stability, we use full-precision (FP32) training. Each batch contains 2 scenes, with a randomly sampled number of input images ranging from 3 to 6 to improve robustness under varying lighting conditions. We evaluate accuracy using the Mean Angular Error (MAE) between the predicted and ground-truth normals. On a single H100 GPU, inference takes approximately 13 seconds for 16 images at a resolution of 512×512, excluding I/O.

Evaluation Dataset

We mainly used three public datasets for evaluation. For quantitative analysis, the following two object-centric datasets were used: **DiLiGenT** (Shi et al. 2016), which contains 10 objects captured under 96 directional lights with orthographic projection and provides 16-bit images, lighting information, masks, and ground truth normals; and **LUCES** (Roberto Mecca and Cipolla 2021), which includes 14 objects imaged under 52 near-field lights using a calibrated perspective camera (12-bit RAW), with full lighting parameters and ground truth normals and depths from 3D scans. Note that we don’t use lighting information.

For the qualitative analysis with more challenging scenes, we use **Multi-illumination dataset** (Murrman et al. 2019). This dataset comprises 1,016 HDR indoor scenes captured under 25 directional lighting conditions using bounced flash illumination. Unlike object-centric datasets with uniform lighting, this scene-level dataset features spatially varying light distributions, where the geometry and materials of each room cause illumination to differ significantly across the scene. Direct lighting is limited to only 7 out of 25 directions; the rest produce purely indirect light via ceiling and wall bounces, yielding realistic, diverse lighting effects across entire environments. No normal map is provided.

We compared our method with SDM-UniPS (Ikehata 2023), Uni-MSPS (Hardy, Quéau, and Tschumperlé 2024), and LINO-UniPS (Li et al. 2025), which were introduced in the related work section.

Evaluation on DiLiGenT and LUCES

We evaluate our method on the DiLiGenT (Shi et al. 2016) (orthographic) and LUCES (Roberto Mecca and Cipolla 2021) (perspective) benchmarks under directional lighting. The purpose of this experiment is to validate that our method performs well on standard universal photometric stereo tasks, and that its effectiveness is not compromised by the introduction of geometric priors. Note that the primary goal is to demonstrate the availability of geometric priors from 3D reconstruction models when multi-illumination cues are unreliable, rather than to achieve the best performance on well-illuminated datasets such as DiLiGenT and LUCES.

Nevertheless, as shown in Table 2 and 3, our method achieves state-of-the-art performance on both datasets. Due to space constraints, all recovered normal maps are provided in the appendix. Notably, incorporating geometric priors significantly improves performance when multi-illumination cues are limited ($K = 1$), enabling robust predictions with fewer input images. However, the method still performs best when sufficient illumination cues are available. For LUCES, while LINO-UniPS achieves competitive accuracy through wavelet-based refinement, our model demonstrates the comparable performance without the wavelet-based refinement.

Interestingly, we observe that using fewer input images sometimes yields better results. We attribute this to training with 3–6 input views, where evaluating on larger sets may lead to performance drops due to distribution shift.

Evaluation on Multi-illumination Dataset

We conducted a qualitative comparison using the Multi-illumination Dataset (Murrman et al. 2019). Seven scenes were randomly selected for the main evaluation, with additional results provided in the Appendix. As ground-truth surface normals are not available for this dataset, our analysis is limited to qualitative assessment; nonetheless, several noteworthy observations emerged.

As shown in Fig. 3, this dataset poses significantly greater challenges than conventional object-centric datasets. Unlike previous datasets with limited lighting directions and simpler scenes, it features complex geometry, greater depth variation, and more realistic spatial layouts. A key finding is that our method clearly benefits from the geometric

priors learned via the pretrained 3D reconstruction model (i.e., VGGT). In particular, it captures plausible normals on floors, walls, and regions with complex depth variation or mirror/transparent surfaces—areas where methods relying solely on multi-illumination cues tend to fail.

Importantly, our method does not depend exclusively on geometric priors. The high-frequency components of the predicted normal maps are comparable to those of other methods, indicating that our approach effectively captures fine surface details from multi-illumination cues. This is especially evident in comparisons with MoGe (Wang et al. 2025b), a monocular depth estimation method, as shown in Fig. 1.

Analytical Studies

Effect of Training Data. To assess the benefit of our training dataset design independently of our architecture, we train SDM-UniPS (Ikehata 2023) from scratch using three different datasets: (1) PS-Mix (Ikehata 2023), rendered under orthographic projection; (2) PS-Perp, our proposed dataset with varying focal lengths under perspective projection; and (3) a hybrid combination of both. These three models are evaluated on 100 newly rendered samples, generated in the same manner as PS-Mix and PS-Perp but with a different focal length. All models were trained for three days using identical hyperparameters.

As shown in Table 1, the model trained solely on PS-Mix performs poorly under strong perspective distortion, with an MAE of 22.18° at 15mm, although performance improves as the focal length increases. In contrast, the model trained on PS-Perp generalizes well to perspective images (e.g., 7.18° at 15mm) but degrades when applied to orthographic images. The hybrid training consistently performs well across all focal lengths, validating the complementary nature of the two datasets.

Effect of Encoder Design. We conduct an analytical study to evaluate the encoder design by comparing three configurations: (1) Encoder_{Geo} only, (2) Encoder_L only, and (3) our full Light-Geometry Dual-Branch Encoder. To isolate encoder effects from decoder influence, we use the single-scale+pixel-embedding decoder from SDM-UniPS (Ikehata 2023). To test generality of our method beyond VGGT (Wang et al. 2025a), we also implement Encoder_{Geo} with MoGe (Wang et al. 2025b), a monocular depth estimator trained on large-scale in-the-wild datasets. All models are trained on the combined PS-Perp and PS-Mix datasets for about 400,000 iterations. Evaluation is conducted on DiLiGenT (Shi et al. 2016) (orthographic) and

Training Dataset	15 mm	35 mm	70 mm	200 mm	Ortho
PS-Mix (Orthographic)	22.18	14.09	10.36	8.89	5.52
PS-Perp (Perspective)	7.18	5.47	5.38	5.75	8.95
PS-Perp + PS-Mix	6.98	5.53	5.53	5.90	5.62

Table 1: Evaluation of SDM-UniPS trained on different training datasets. MAE ↓ under varying focal lengths (mm).

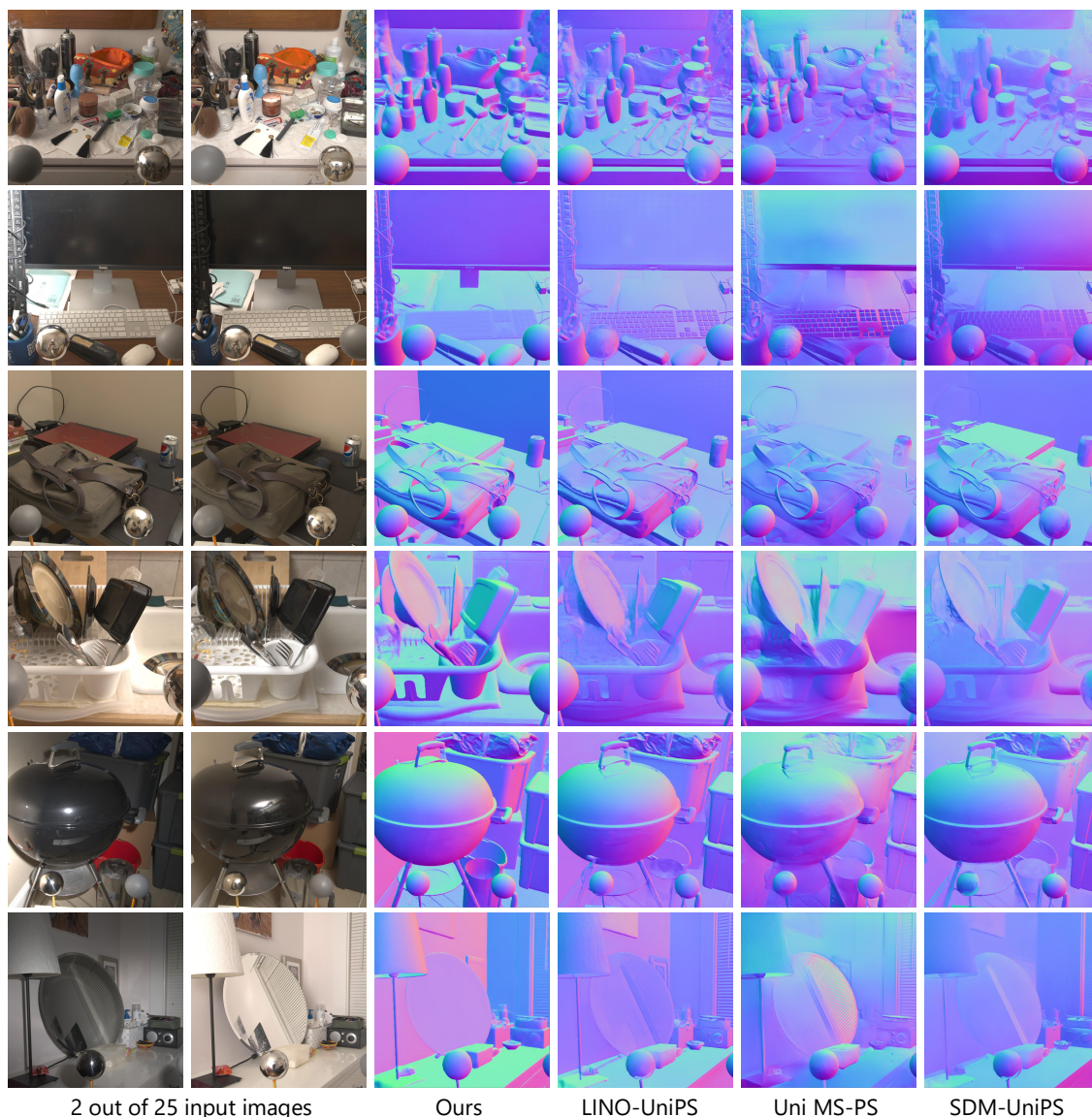


Figure 3: Qualitative comparison on Multi-illumination Dataset (Murmann et al. 2019).

LUCES (Roberto Mecca and Cipolla 2021) (perspective) benchmarks using MAE over all objects.

As shown in Table 4, EncoderIL only degrades significantly when multi-illumination cues are absent ($K = 1$), resulting in poor normal estimation. With larger K , these cues become more reliable, improving performance. In contrast, EncoderGeo only does not scale well with more input images, despite using illumination cues in the decoder, highlighting the importance of extracting them in the encoder. Our Dual-Branch Encoder, while slightly worse than Encoder_{IL} only on DiLiGenT with $K = 4$ and $K = 16$, performs robustly across both $K = 1$ and $K = 16$. It notably outperforms others on LUCES, likely because geometric priors are more effective under perspective projection. Comparing backbones, VGGT consistently outperforms MoGe, validating our choice. However, both show similar trends,

confirming that geometric priors benefit PS tasks regardless of the pretrained model.

Comparison with 3D Scan. While the Multi-illumination Dataset highlights the advantage of our method under limited lighting diversity, its lack of ground-truth normals and single-directional lighting may raise concerns. To address this, we conduct a quantitative evaluation using four images captured under weak, non-uniform lighting from a selfie ring light with ambient illumination. Lighting movement was minimized to suppress cues, and object masks were not provided. This setup is intentionally difficult and serves as an analytical stress test rather than a realistic scenario.

3D scans were obtained using an EinScan-SE scanner and aligned for evaluation. As shown in Fig. 4 the results, both SDM-UniPS and LINO-UniPS fail almost completely,

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
SDM-UniPS (Ikehata 2023)	1.45	3.50	7.54	5.19	4.48	7.69	10.76	4.59	4.41	8.43	5.80
Uni MS-PS (Hardy et al. 2024)	1.93	3.56	6.53	4.13	4.12	7.35	9.88	4.31	4.49	7.45	5.38
LINO-UniPS (Li et al. 2025)	1.77	2.62	6.22	3.38	4.38	5.14	8.60	4.07	4.54	6.75	4.75
Ours	2.63	2.46	5.95	3.27	3.93	5.00	8.54	3.81	4.00	6.88	4.65
Ours (K=16)	2.75	2.45	6.04	3.44	3.84	5.00	8.88	3.96	4.24	7.12	4.77
Ours (K=4)	3.49	2.73	7.66	5.58	4.09	5.63	11.15	4.62	5.43	8.76	5.91
Ours (K=1)	6.70	5.74	17.26	12.94	10.78	10.26	26.53	9.72	10.06	18.60	12.86

Table 2: Evaluation on DiLiGenT (Mean Angular Errors in degrees). All 96 images were used except where K is shown.

Method	Ball	Bell	Bowl	Buddha	Bunny	Cup	Die	Hippo	House	Jar	Owl	Queen	Squirrel	Tool	Avg
SDM-UniPS	11.77	12.92	8.66	18.16	8.83	11.36	7.22	8.95	25.91	8.84	12.82	15.30	15.92	12.58	12.80
Uni MS-PS	11.62	11.66	7.96	13.38	10.02	7.92	6.50	8.80	25.62	6.35	12.07	12.77	12.18	11.24	11.29
LINO-UniPS	9.65	8.97	8.26	13.30	5.67	8.30	6.25	5.82	22.69	6.13	9.29	9.98	10.56	7.55	9.46
Ours	7.59	10.22	8.13	13.11	5.50	10.05	3.79	5.62	21.84	6.17	10.76	10.56	9.86	8.61	9.42
Ours (K=16)	7.42	10.71	7.83	13.10	5.35	9.80	3.71	5.63	22.00	6.43	10.56	10.79	11.17	8.65	9.62
Ours (K=4)	7.75	12.22	7.46	14.25	5.39	8.57	4.06	6.15	24.14	6.85	12.29	12.85	13.27	8.52	10.27
Ours (K=1)	9.08	10.12	11.78	15.71	11.14	15.05	7.31	12.31	32.71	8.80	16.11	20.67	21.60	10.78	14.33

Table 3: Evaluation on LUCES (Mean Angular Errors in degrees). All 52 images were used except where K is shown.

Encoder	DiLiGenT			LUCES		
	K=1	K=4	K=16	K=1	K=4	K=16
Encoder _{Geo} w/ MoGe	13.63	8.96	8.33	15.50	12.32	11.01
Encoder _{Geo} w/ VGGT	13.07	8.27	6.90	14.70	11.90	11.12
Encoder _{TL}	19.03	6.50	4.96	19.82	11.40	10.36
Dual-Branch Encoder	12.84	6.81	5.19	14.82	10.60	9.82

Table 4: Ablation study on encoder configurations. MAE (\downarrow) for different K on DiLiGenT (Shi et al. 2016) and LUCES (Roberto Mecca and Cipolla 2021). All models use a single decoder.

while our method, though not perfect, performs significantly better. This suggests that leveraging pretrained geometric knowledge is highly effective, especially under conditions where traditional photometric stereo struggles.

Conclusion

We introduced GeoUniPS, a geometry-aware universal photometric stereo network designed to address the limitations of existing methods under insufficient or biased illumination. Motivated by the observation that real-world scenes often lack diverse lighting cues, we proposed leveraging high-level geometric priors from pretrained 3D reconstruction models. To this end, we designed a Light-Geometry Dual-Branch Encoder that jointly captures lighting-aware features and lighting-invariant geometric features. GeoUniPS achieves state-of-the-art performance on both orthographic and perspective benchmarks and demonstrates strong qualitative results on complex real-world scenes.

While we leverage 3D reconstruction models as geometric priors, exploring alternative backbones such as diffusion-based geometry estimators or other large-scale pretrained

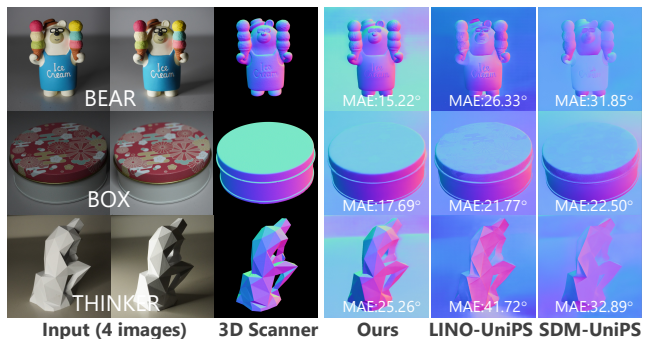


Figure 4: Results for scenes under limited lighting cues, excluding object masks.

models (e.g., segmentation models), remains future work. Additionally, the performance boost under sufficient lighting cues is limited, likely due to our simple feature concatenation. Designing better fusion strategies is an important direction ahead.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24K02966 and DENSO IT LAB Recognition, Control and Learning Algorithm Collaborative Research Chair (Science Tokyo).

References

- Alldrin, N.; and Kriegman, D. 2007. Toward Reconstructing Surfaces With Arbitrary Isotropic Reflectance: A Stratified Photometric Stereo Approach. In *ICCV*.
- Chen, G.; Han, K.; Shi, B.; Matsushita, Y.; and Wong, K. K. 2019. Self-Calibrating Deep Photometric Stereo Net-

- works. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8731–8739.
- Chen, G.; Han, K.; Shi, B.; Matsushita, Y.; and Wong, K.-Y. K. 2019. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8739–8747.
- Chen, G.; Han, K.; and Wong, K.-Y. K. 2018. PS-FCN: A Flexible Learning Framework for Photometric Stereo. *ECCV*.
- Chen, G.; Waechter, M.; Shi, B.; Wong, K.-Y. K.; and Matsushita, Y. 2020. What is learned in deep uncalibrated photometric stereo? In *European Conference on Computer Vision*, 745–762. Springer.
- Goldman, D. B.; Curless, B.; Hertzmann, A.; and Seitz, S. M. 2010. Shape and Spatially-Varying BRDFs from Photometric Stereo. *IEEE TPAMI*, 32(6): 1060–1071.
- Hardy, C.; Quéau, Y.; and Tschumperlé, D. 2024. Uni MS-PS: A multi-scale encoder-decoder transformer for universal photometric stereo. *Computer Vision and Image Understanding*, 248: 104093.
- Ikehata, S. 2018. CNN-PS: CNN-based Photometric Stereo for General Non-convex Surfaces. In *ECCV*.
- Ikehata, S. 2021. PS-Transformer: Learning Sparse Photometric Stereo Network using Self-Attention Mechanism. In *BMVC*.
- Ikehata, S. 2022. Universal Photometric Stereo Network using Global Lighting Contexts. In *CVPR*.
- Ikehata, S. 2023. Scalable, Detailed and Mask-free Universal Photometric Stereo. In *CVPR*.
- Ikehata, S.; and Asano, Y. 2024. Physics-Free Spectrally Multiplexed Photometric Stereo under Unknown Spectral Composition. In *ECCV*.
- Ikehata, S.; Wipf, D.; Matsushita, Y.; and Aizawa, K. 2012. Robust photometric stereo using sparse regression. In *CVPR*.
- Kaya, B.; Kumar, S.; Oliveira, C.; Ferrari, V.; and Van Gool, L. 2021. Uncalibrated Neural Inverse Rendering for Photometric Stereo of General Surfaces. 3804–3814.
- Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. In *proc. ICLR*.
- Lee, J.; Lee, Y.; Kim, J.; Kosiorok, A.; Choi, S.; and Teh, Y. W. 2019. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *ICML*, 3744–3753.
- Leroy, V.; Cabon, Y.; and Revaud, J. 2024. Grounding Image Matching in 3D with MAST3R.
- Li, H.; Chen, H.; Ye, C.; Chen, Z.; Li, B.; Xu, S.; Guo, X.; Liu, X.; Wang, Y.; Zhang, B.; et al. 2025. Light of Normals: Unified Feature Representation for Universal Photometric Stereo. *arXiv preprint arXiv:2506.18882*.
- Li, J.; and Li, H. 2022. Neural Reflectance for Shape Recovery With Shadow Handling. In *CVPR*.
- Murmann, L.; Gharbi, M.; Aittala, M.; and Durand, F. 2019. A Multi-Illumination Dataset of Indoor Object Appearance. In *ICCV*.
- Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision Transformers for Dense Prediction. In *ICCV*.
- Roberto Mecca, I. B., Fotios Logothetis; and Cipolla, R. 2021. LUCES: A Dataset for Near-Field Point Light Source Photometric Stereo. In *BMVC*.
- Sarno, F.; Kumar, S.; Kaya, B.; Huang, Z.; Ferrari, V.; and Van Gool, L. 2022. Neural Architecture Search for Efficient Uncalibrated Deep Photometric Stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 361–371.
- Shi, B.; Tan, P.; Matsushita, Y.; and Ikeuchi, K. 2014. Bi-polynomial Modeling of Low-frequency Reflectances. *IEEE TPAMI*, 36(6): 1078–1091.
- Shi, B.; Wu, Z.; Mo, Z.; Duan, D.; Yeung, S.-K.; and Tan, P. 2016. A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3707–3716.
- Tang, Z.; Fan, Y.; Wang, D.; Xu, H.; Ranjan, R.; Schwing, A.; and Yan, Z. 2025. MV-DUST3R+: Single-Stage Scene Reconstruction from Sparse Views In 2 Seconds. In *CVPR*.
- Tani, T.; and Maehara, T. 2018. Neural inverse rendering for general reflectance photometric stereo. In *International Conference on Machine Learning*, 4857–4866. PMLR.
- Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Ruppel, C.; and Novotny, D. 2025a. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.
- Wang, R.; Xu, S.; Dai, C.; Xiang, J.; Deng, Y.; Tong, X.; and Yang, J. 2025b. MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision. In *CVPR*.
- Wang, R.; Xu, S.; Dong, Y.; Deng, Y.; Xiang, J.; Lv, Z.; Sun, G.; Tong, X.; and Yang, J. 2025c. MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details. *arXiv:2507.02546*.
- Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.
- Woodham, R. J. 1980. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1): 139–144.
- Wu, L.; Ganesh, A.; Shi, B.; Matsushita, Y.; Wang, Y.; and Ma, Y. 2010. Robust Photometric Stereo via Low-Rank Matrix Completion and Recovery. In *ACCV*.
- Yamaguchi, M.; Shibata, T.; Yachida, S.; Yokoyama, K.; and Hosoi, T. 2025. MDCN-PS: Monocular-Depth-Guided Coarse Normal Attention for Robust Photometric Stereo. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 3342–3351.
- Yao, Z.; Li, K.; Fu, Y.; Hu, H.; and Shi, B. 2020. GPS-Net: Graph-based Photometric Stereo Network. *NeurIPS*.