MVCPS-NeuS: Multi-view Constrained Photometric Stereo for Neural Surface Reconstruction

Hiroaki Santo Fumio Okura Yasuyuki Matsushita
Graduate School of Information Science and Technology, Osaka University
{santo.hiroaki, okura, yasumat}@ist.osaka-u.ac.jp

Abstract

Multi-view photometric stereo (MVPS) recovers a high-fidelity 3D shape of a scene by benefiting from both multi-view stereo and photometric stereo. While photometric stereo boosts detailed shape reconstruction, it necessitates recording images under various light conditions for each viewpoint. In particular, calibrating the light directions for each view significantly increases the cost of acquiring images. To make MVPS more accessible, we introduce a practical and easy-to-implement setup, multi-view constrained photometric stereo (MVCPS), where the light directions are **unknown but constrained** *to move together with the camera. Unlike conventional multi-view uncalibrated photometric stereo, our constrained setting reduces the ambiguities of surface normal estimates from per-view linear ambiguities to a single and global linear one, thereby simplifying the disambiguation process. The proposed method integrates the ambiguous surface normal into neural surface reconstruction (NeuS) to simultaneously resolve the global ambiguity and estimate the detailed 3D shape. Experiments demonstrate that our method estimates accurate shapes under sparse viewpoints using only a few multi-view constrained light sources.*

1. Introduction

Multi-view photometric stereo (MVPS) is a 3D reconstruction approach that combines photometric stereo (PS) and multi-view stereo (MVS). In a conventional setting, a scene is recorded from multiple viewpoints, and at each viewpoint, multiple images are captured under varying light directions [14, 19, 29, 40]. With the advancement of neural surface reconstruction [37, 41], recent works [14, 40] have achieved high-fidelity 3D reconstruction by incorporating the surface normal derived from PS via inverse rendering.

To eliminate the cost of calibrating the cameras and light directions in MVPS, it is generally favored to work in an *uncalibrated* setting, where the camera can freely move and light directions are treated unknown. For determining the camera parameters including its postures, we can safely rely on mature Structure-from-Motion (SfM) methods; however, uncalibrated light directions pose a problem in determining surface details. Namely, when the light directions are unknown, the problem becomes *uncalibrated* photometric stereo (UPS), and it is known that surface normal can only be estimated up to a linear ambiguity. As a result, in uncalibrated MVPS, it yields an ambiguous surface normal map for each view, where each of them has a different liner ambiguity. Due to the per-view linear ambiguity in the estimated surface normal maps, it remains challenging to fully take advantage of PS's capability in uncalibrated MVPS.

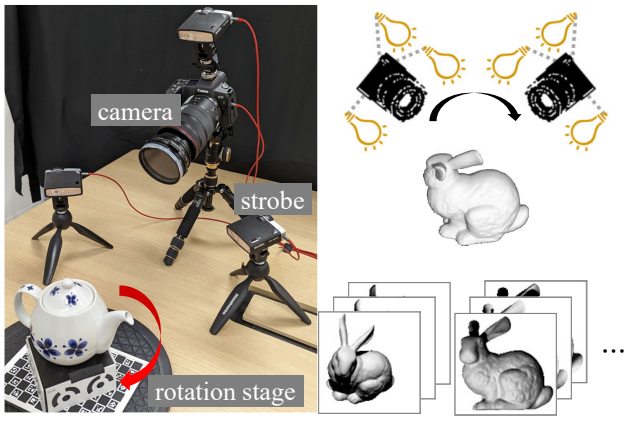

Figure 1. The minimal setting for multi-view *constrained* photometric stereo, equipping a camera, three strobes, and a manual rotation stage with markers. The camera and light sources move together with respect to the target object.

In this paper, we consider a setting of MVPS, where light directions are unknown but constrained to move together with a camera, which we call multi-view **constrained** photometric stereo (MVCPS). Such a setup can be easily achieved by employing a rig that secures a camera and light sources, or by using a rotation stage to move a target object in front of a camera and light sources, as shown in Fig. 1. In fact, the setting is hardly new, and most existing single-view or multi-view PS works employ such setups [19, 22, 25, 27, 31, 33–

35, 45] because of the ease of implementation.

We show that, in MVCPS with unknown light directions, the ambiguity of surface normal can be reduced from per-view ones to a single and global one. By jointly factorizing the multi-view & multi-light observations, our method derives multi-view surface normal maps that have a unique and common linear ambiguity. Leveraging the surface normal maps with reduced ambiguity, we develop a neural surface reconstruction method that jointly resolves the ambiguity, thereby achieving high-fidelity 3D shape recovery. The proposed method also introduces a confidence estimation for surface normal based on the reduced ambiguity, enhancing robustness against outliers like shadows.

Our experimental results demonstrate that the proposed method yields more accurate estimations under minimal light directions and sparse views, *e.g*., three light sources and four viewpoints, compared to state-of-the-art multi-view uncalibrated photometric stereo methods. To sum up, this paper provides the following three contributions:

- We show that our MVCPS reduces the per-view ambiguities into a single and global ambiguity, which allows for better disambiguation and shape recovery.
- We develop an outlier detection method based on the global ambiguity for robust estimation.
- The proposed method achieves detailed shape reconstruction by integrating the ambiguous surface normal into neural surface reconstruction.

2. Related Work

In this section, we first review previous works of single- and multi-view PS. We then describe recent works of neural surface representations.

Single-view Photometric Stereo Conventional PS [36, 38] started with the Lambertian assumption [17] and known light directions. The assumption of known light directions has been later relaxed to deal with *uncalibrated* settings, where light directions are treated unknown.

UPS simultaneously estimates scene shape and light directions and eliminates the necessity of light source calibration, while there exists a linear ambiguity. Early works [1, 9] employ the Lambertian assumption and factorize the input observations into surface normals and light directions. To resolve the ambiguity, the uses of known surface albedo [9], shadows [16], and specularity [4, 5] have been explored.

Recent works [2] use a learning-based approach to resolve the ambiguity by data prior. As discussed in [3], the learning-based method also implicitly uses shadows and specularity to disambiguate the estimation. More recently, inverse rendering-based approaches [11, 18, 20] have been proposed. While they achieve accuracy comparable to that of calibrated settings in scenes with dense light sources, they still face challenges when the light sources are sparse.

Multi-view Photometric Stereo (MVPS) Early works of MVPS start with estimating a base shape from multi-view observations, such as a mesh [29], SDF volume [22], and sparse point cloud [19], and then refine the shape using PS to achieve detailed shape recovery. Kaya *et al*. [13] propose an integration of recent deep learning-based MVS and PS. Subsequently, their method has been extended [12, 14] to account for the uncertainties in the depth and normal estimations. While these methods achieve good 3D reconstruction, they assume dense observations, *i.e*., the number of viewpoints and light sources is large enough to obtain accurate estimations from both MVS and PS.

Along with the recent advances in neural surface reconstruction, Yang *et al*. [40] propose PS-NeRF, which incorporates normal maps estimated by PS into neural surface reconstruction. They estimate per-view normal maps in advance and use them to optimize a neural surface. They use the state-of-the-art learning-based UPS method [2] for normal map estimation for each view. However, when only a limited number of light sources are available, UPS fails to accurately estimate the normal maps, resulting in a collapsed shape estimation.

More recently, MVPSNet [44] proposes a feature extractor to leverage shading information observed under varying lights, thereby facilitating improved stereo matching. They also use the UPS method [2] to estimate the light directions.

The proposed method shares the spirit of PS-NeRF. However, to overcome the challenges posed by limited viewpoints and light sources, we propose to simultaneously optimize shapes and disambiguate surface normal in neural surface reconstruction instead of involving per-view disambiguation of the surface normal.

Neural Surface Reconstruction Neural implicit surfaces have achieved remarkable advancements in novel view synthesis [26] and shape recovery [28, 37, 41, 42]. While these methods achieve high-quality shape reconstruction, they require a large number of images to accurately optimize the neural surface.Scenes with sparse viewpoints have been challenging, and several works tackle this problem. Long *et al*. [23] propose a pre-trained encoder that estimates a coarse shape from sparse inputs. In the direction of using prior from a pre-trained model, Wu *et al*. [39] introduce the use of multi-view consistency, which has been used in conventional MVS [7]. Yu *et al*. [43] propose the more explicit use of a pre-trained prior, *i.e*., a monocular depth and normal estimation method. They estimate the per-view depth and normal maps from a single RGB image by [6] for supervision.

Using prior knowledge from a large amount of data is effective in cases where the training data covers test scenes. In contrast, the proposed method uses additional observations under a few different light sources and demonstrates superior performance in scenes with sparse viewpoints.

3. Proposed Method

Our method observes a target object from v viewpoints under l light directions and obtains observations $\{\mathbf{M}_1, \cdots, \mathbf{M}_v\} \subset \mathbb{R}^{p \times l}$, where p represents the number of pixels. We assume that cameras' intrinsics and extrinsics are known by SfM or markers placed in the scene. We further assume that foreground masks $\{\mathbf{s}_1, \cdots, \mathbf{s}_v\} \subset \mathbb{R}^p$ are available. The proposed method jointly decomposes the observations $\{\mathbf{M}_i\}$ into per-view surface normal maps and shared light directions w.r.t. the camera. Subsequently, a neural surface is optimized with the supervision of color observations, foreground masks, and the decomposed surface normals. In this section, we describe these details.

3.1. Multi-view Constrained Photometric Stereo (MVCPS)

We begin with a naive extension of factorization-based UPS for multi-view observations. We then show that, by assuming the camera and lights move together, we can jointly factorize the multi-view observations to reduce the ambiguity of surface normal maps.

Singular Value Decomposition-based multi-view UPS Hayakawa [9] propose a UPS method that decomposes single view observations under varying lights using singular value decomposition (SVD). It can be simply extended to the multi-view context by solving UPS for each view. The observation at i-th view, $\mathbf{M}_i \in \mathbb{R}^{p \times l}$, can be decomposed into left- and right-singular vectors $\mathbf{U}_i \in \mathbb{R}^{p \times p}$, $\mathbf{V}_i \in \mathbb{R}^{l \times l}$ and singular values $\mathbf{\Sigma}_i \in \mathbb{R}^{p \times l}$ as:

$$\mathbf{M}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^\top.$$

$\mathbf{U}_i^\top \mathbf{U}_i = \mathbf{I}_p$, $\mathbf{V}_i^\top \mathbf{V}_i = \mathbf{I}_l$, the diagonal elements of Σ_i are singular values, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. With the Lambertian assumption, it becomes $\operatorname{rank} \mathbf{M}_i = 3$ for $l \geq 3$ under different and non-coplanar light directions, the observations $\mathbf{M}$ can be written as a product of low-dimensional matrices as

$$\mathbf{M}_i = \underbrace{\mathbf{U}_i' \mathbf{\Sigma}_i'}_{\hat{\mathbf{N}}_i} \underbrace{\mathbf{V}_i'^\top}_{\hat{\mathbf{L}}_i^\top}, \tag{1}$$

where $\mathbf{U}_i' \in \mathbb{R}^{p \times 3}$ and $\mathbf{V}_i' \in \mathbb{R}^{l \times 3}$ are first three singular vectors of $\mathbf{U}_i$ and $\mathbf{V}_i$, respectively, and $\mathbf{\Sigma}_i'$ is a 3×3 diagonal matrix containing singular values. The decomposed $\hat{\mathbf{N}}_i \in \mathbb{R}^{p \times 3}$ and $\hat{\mathbf{L}}_i \in \mathbb{R}^{l \times 3}$ denote the estimates of the surface normals and light directions for i-th view, respectively.

The solution is known to contain a linear ambiguity $\mathbf{X}_i \in \mathbb{R}^{3 \times 3}$ (where $\det(\mathbf{X}_i) \neq 0$) for each view, which means any invertible 3×3 matrix $\mathbf{X}_i$ can be inserted as

$$\mathbf{M}_i = \left(\hat{\mathbf{N}}_i \mathbf{X}_i\right)\left(\mathbf{X}_i^{-1} \hat{\mathbf{L}}_i^\top\right) = \hat{\mathbf{N}}_i' \hat{\mathbf{L}}_i' \tag{2}$$

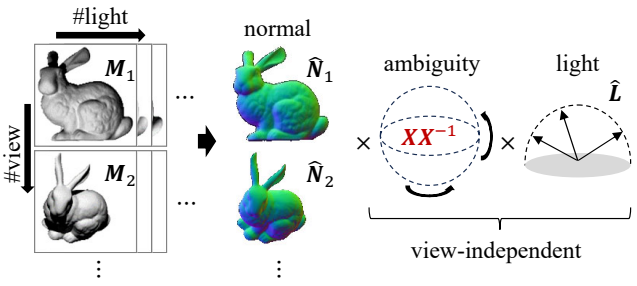

Figure 2. Decomposition of multi-view observations by HO-GSVD

to obtain different combinations of surface normals and light directions. As such, a straightforward application of UPS to the multi-view setting suffers from *per-view* ambiguities $\{\mathbf{X}_i\}$.

Our MVCPS via Higher-Order Generalized SVD Different from the general uncalibrated case described above, our MVCPS setting assumes that the camera and lights move together. In this setting, we can assume unknown but consistent light directions $\hat{\mathbf{L}}_i$ across all views with respect to the camera as

$$\forall i, \hat{\mathbf{L}}_i = \hat{\mathbf{L}}. \tag{3}$$

With the knowledge of consistent light directions $\hat{\mathbf{L}}$, we cast the problem of multi-view UPS to matrix factorization based on higher-order generalized SVD (HO-GSVD) [15, 30]. With the HO-GSVD, we can decompose an arbitrary number of observations $\{\mathbf{M}_i\}$ into $\{\mathbf{U}_i\}$, $\{\mathbf{\Sigma}_i\}$, and common $\mathbf{V} \in \mathbb{R}^{l \times l}$ as

$$\mathbf{M}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}^\top. \tag{4}$$

Following Eq. (1), we obtain estimates of surface normal $\{\hat{\mathbf{N}}_i\}$ and a consistent light direction $\hat{\mathbf{L}}$:

$$\mathbf{M}_i = \hat{\mathbf{N}}_i \mathbf{X} \mathbf{X}^{-1} \hat{\mathbf{L}}^\top, \tag{5}$$

where $\hat{\mathbf{N}}_i = \mathbf{U}_i' \mathbf{\Sigma}_i'$ and $\hat{\mathbf{L}} = \mathbf{V}' \in \mathbb{R}^{l \times 3}$ is the first three singular vectors of $\mathbf{V}$. The matrix $\mathbf{X} \in \mathbb{R}^{3 \times 3}$ $(\det(\mathbf{X}) \neq 0)$ is the global linear ambiguity. Compared to the per-view linear ambiguities $\{\mathbf{X}_i\}$ that appeared in Eq. (2), with our method, we can reduce the ambiguity to a global one $\mathbf{X}$ shared by all the views (Fig. 2). While the decomposition of observations $\{\mathbf{M}_i\}$ into $\{\hat{\mathbf{N}}_i\}$ and $\hat{\mathbf{L}}$ can also be achieved by applying SVD to the concatenated observation matrices $\left[\mathbf{M}_1^\top, \mathbf{M}_2^\top, \cdots\right]^\top$, HO-GSVD offers a more accurate low-rank approximation (Eq. (5)) when more than three light sources are available. As we will see in the next section, the proposed method resolves the ambiguity $\mathbf{X}$ through neural surface reconstruction using the supervision by the decomposed per-view surface normals $\{\hat{\mathbf{N}}_i\}$.

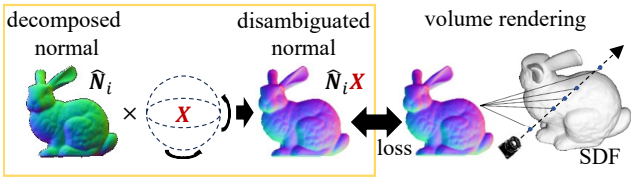

Figure 3. Optimization of SDF by decomposed surface normal

3.2. Neural Surface Reconstruction for MVCPS

Once we have per-view surface normals $\{\hat{\mathbf{N}}_i\}$ with ambiguity $\mathbf{X}$, we fuse them to a neural surface reconstruction method by extending NeuS [37] to recover the detailed 3D shape and resolve the ambiguity $\mathbf{X}$ simultaneously (Fig. 3).

Following the work of NeuS [37], we represent a scene surface $\mathbf{S}$ by a zero-level set of a neural signed distance function (SDF) $f_\theta : \mathbb{R}^3 \to \mathbb{R}$ as

$$\mathbf{S} = \{\mathbf{x} \in \mathbb{R}^3 | f_\theta(\mathbf{x}) = 0\},$$

where $\mathbf{x}$ is a 3D point, and θ is a learnable parameter. Our MVCPS provides the surface normal maps $\{\hat{\mathbf{N}}_i\}$ with an unknown global ambiguity represented as $\mathbf{X}$. We will call $\{\hat{\mathbf{N}}_i\mathbf{X}\}$ as disambiguated surface normal maps hereafter.

Normal consistency loss The gradient of the SDF, $\nabla f_\theta(\mathbf{x})$ at a point $\mathbf{x}$, represents the surface normal of the point. By applying volume rendering along a camera ray, $\mathbf{p}(t) = \mathbf{o} + t\mathbf{v}$ $(t > 0)$, where $\mathbf{o} \in \mathbb{R}^3$ and $\mathbf{v} \in \mathbb{S}^2$ represent a camera position and viewing direction, respectively, we can compute the surface normal $\mathbf{n}^*$ from the SDF as

$$\mathbf{n}^*(\mathbf{o}, \mathbf{v}) = \int_0^\infty w(t) \nabla f_\theta(\mathbf{p}(t))\, dt.$$

The weight $w(t)$ for a point on the ray is computed from the volume density $\sigma(t)$ as

$$w(t) = T(t)\sigma(t),$$

where $T(t) = \exp\left(-\int_o^t \sigma(u)du\right)$ is an accumulated transmittance. We employ the normalized S-density [37] as the volume density $\sigma(t)$ computed from the SDF f_θ. We also follow the uniform and near-surface sampling strategy in [37] for computing the integration.

To optimize the ambiguity matrix $\mathbf{X} \in \mathbb{R}^{3\times3}$ through backpropagation, we introduce a normal consistency loss between the SDF surface normal $\mathbf{n}^*$ and the disambiguated surface normal $\{\hat{\mathbf{N}}_i\mathbf{X}\}$:

$$\mathcal{L}_{\text{normal}} = \sum_{(\mathbf{o},\mathbf{v})\in\chi} ||\tau_c(\mathbf{n}^*(\mathbf{o}, \mathbf{v})) - \hat{\mathbf{n}}(\mathbf{o}, \mathbf{v})||_1, \quad (6)$$

where τ_c denotes the transformation of a surface normal from the world coordinates to camera coordinates, and $\hat{\mathbf{n}}(\mathbf{o}, \mathbf{v})$ is the surface normal derived from the corresponding pixel of disambiguated surface normals $\{\hat{\mathbf{N}}_i\mathbf{X}\}$.

Although the normal loss $\mathcal{L}_{\text{normal}}$ does not resolve the ambiguities by itself, the SDF surface normals $n^*(\mathbf{o}, \mathbf{v})$ are also constrained by photo and mask consistency losses in the case of NeuS, which constrains the ambiguity matrix $\mathbf{X}$. Concurrently, the SDF is refined using the disambiguated surface normals, which contain detailed shape information derived from MVCPS.

While photo-consistency loss is only applicable when the same scene point is observed from multiple viewpoints, the normal loss can constrain the surface orientation even when the scene point is observed only from a single view, which enables the recovery of shape from sparse observations.

Other loss functions We follow the original NeuS implementation for the remaining losses, considering color, mask, and regularization. We here briefly recap them.

To optimize SDF f_θ by color observations, we follow the volume rendering of a radiance field $\mathbf{c}_\phi : \mathbb{R}^3 \times \mathbb{S}^2 \to \mathbb{R}^3_+$ proposed by NeRF [26] and render the color $\mathbf{C}^* \in \mathbb{R}^3_+$ as

$$\mathbf{C}^*(\mathbf{o}, \mathbf{v}) = \int_0^\infty w(t)\mathbf{c}_\phi(\mathbf{p}(t), \mathbf{v})\, dt.$$

The color loss is computed between rendered and observed intensities as

$$\mathcal{L}_{\text{color}} = \sum_{(\mathbf{o},\mathbf{v})\in\chi} ||\mathbf{C}^*(\mathbf{o}, \mathbf{v}) - \mathbf{C}(\mathbf{o}, \mathbf{v})||_1,$$

where χ is the set of sampled rays and $\mathbf{C}(\mathbf{o}, \mathbf{v})$ is the observed intensity of the corresponding pixel in the input images. We employ the L1 loss for robust estimation. For the observed intensities $\mathbf{C}$, we use the median image of input images captured under different light directions, as following [19], or images captured under natural illumination when available. We also employ the mask loss and Eikonal loss [8] as described in [37]. The mask loss $\mathcal{L}_{\text{mask}} = \text{BCE}(s^*(\mathbf{o}, \mathbf{v}), s(\mathbf{o}, \mathbf{v}))$ is defined as the binary cross entropy between the rendered mask $s^*(\mathbf{o}, \mathbf{v})$ and input mask $s(\mathbf{o}, \mathbf{v})$. The Eikonal loss regularizes the norm of the gradients of the SDF f_θ as

$$\mathcal{L}_{\text{reg}} = \sum_{\mathbf{x}\in\psi} \left(||\nabla f_\theta(\mathbf{x})||_2 - 1\right)^2, \quad (7)$$

where ψ is the set of sampled points on the rays.

Training The SDF f_θ and radiance field $\mathbf{c}_\phi$ are represented by multilayer perceptrons (MLPs). We optimize the parameters of MLPs, θ and ϕ, and the ambiguity matrix $\mathbf{X}$ using the following overall loss function:

$$\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \quad (8)$$

where λ_{mask}, λ_{normal}, and λ_{reg} are weights for the losses. To achieve better convergence in the simultaneous optimization of the SDF parameters and ambiguity matrix $\mathbf{X}$, we use a coarse-to-fine optimization strategy with positional encoding [21], which gradually enables high-frequency encoding as the training progresses.

3.3. Confidence Estimation

The proposed factorization method in MVCPS assumes a Lambertian surface without shadows. However, in the real world, outliers like shadows and specularity cannot be neglected, as they lead to estimation errors. Particularly with a limited number of light sources, *e.g.*, only three, accurate estimation becomes difficult even using robust optimization algorithms or recent learning-based methods. In the proposed method, inspired by the fact that the ambiguity matrix $\mathbf{X}$ is shared across all views, we introduce a confidence estimation during training to migrate adverse effects arising from inaccurate estimations.

Let us represent the loss function $\mathcal{L}$ as a function of the ambiguity matrix $\mathbf{X}$ and the view index i where rays are sampled, denoted as $\mathcal{L}(\mathbf{X}, i)$. We consider two different views, u and v, and assess the changes in the loss at u-th view by applying the gradient of the loss at v-th view. The changes $\Delta\mathcal{L}$ is written as

$$\Delta\mathcal{L}_{u,v} = \mathcal{L}(\mathbf{X}, u) - \mathcal{L}\left(\mathbf{X} - \alpha\frac{\partial\mathcal{L}(\mathbf{X}, v)}{\partial\mathbf{X}}, u\right), \quad (9)$$

where α is the step size of the updates. When assuming that the gradient at the v-th view, $\frac{\partial\mathcal{L}(\mathbf{X},v)}{\partial\mathbf{X}}$, improves the ambiguity matrix $\mathbf{X}$, it is expected that the loss at u-th view should improve, given that the ambiguity matrix $\mathbf{X}$ is shared across all views. Nonetheless, if the decomposed surface normal is incorrect at the u-th view, for example, due to shadows, the ambiguity matrix $\mathbf{X}$ becomes irrelevant for such pixels. As a result, the losses at these pixels may increase or decrease, regardless of the improvements in $\mathbf{X}$.

Building on this premise, we propose a variance-based confidence estimation method. We compute the changes of the loss $\Delta\mathcal{L}_{u,v}$ for each pixel over different training steps and construct the variance maps $\{\mathbf{q}_1, \cdots, \mathbf{q}_v\} \subset \mathbb{R}_+^p$. Although it is not guaranteed that the gradient at the v-th view is always accurate, from a statistical standpoint, the ambiguity matrix $\mathbf{X}$ is expected to improve as training progresses. Therefore, by statistically analyzing the changes in loss across different training steps, we expect the surface normals of pixels exhibiting lower variance to have higher confidence. From the variance maps $\{\mathbf{q}_i\}$, we compute the confidence $\{\bar{\mathbf{q}}_i\} \in [0, 1]^p$ by applying exponential mapping,

$$\bar{\mathbf{q}}_i = \exp\left(-\kappa\frac{\mathbf{q}_i}{\max(\mathbf{q}_i)}\right), \quad (10)$$

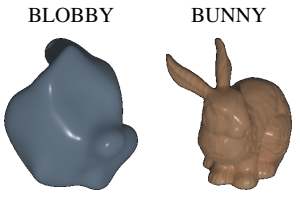

Figure 4. Our synthetic dataset

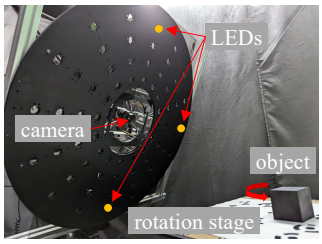

Figure 5. Setup of our real-world experiment

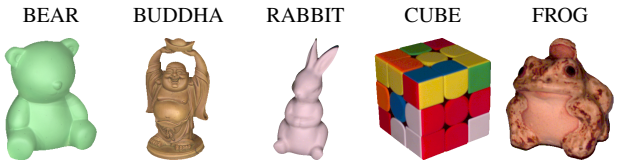

Figure 6. Scenes for our real-world experiments

where κ is a hyperparameter to define the confidence for a pixel with maximum variance, *i.e.*, $\frac{\mathbf{q}_i}{\max(\mathbf{q}_i)} = 1$. In our implementation, the parameter κ is set to 5. The normal loss $\mathcal{L}_{\text{normal}}$ is weighted by the confidence of each pixel.

4. Experiments

We evaluate the proposed method on our synthetic dataset, a public MVPS real-world dataset, DiLiGenT-MV, and our real-world dataset. In the following sections, we detail the experimental settings and present the evaluation results.

Comparison methods We compare the proposed method with NeuS [37], PS-NeRF [40], and MonoSDF [43]. NeuS and MonoSDF are MVS methods based on neural surface reconstruction, taking RGB images as inputs. As RGB images, we use median images of images captured with varying light directions. PS-NeRF is a state-of-the-art multi-view UPS that optimizes the neural surface using RGB images and per-view normal maps, which are estimated from single-view images captured under multiple unknown lighting directions. For a fair comparison, we adopt the same neural surface representation as proposed by NeuS across all methods but use the respective loss functions. More specifically, when training the comparison methods, we replace the normal loss of Eq. (6) by:

$$\mathcal{L}_{\text{normal}} = \sum_{(\mathbf{o},\mathbf{v})\in\chi} ||\tau_{\text{c}}\left(\mathbf{n}^*(\mathbf{o},\mathbf{v})\right) - \hat{\mathbf{n}}'(\mathbf{o},\mathbf{v})||_1, \quad (11)$$

where surface normal $\hat{\mathbf{n}}'$ is computed in the individual comparison methods. For PS-NeRF, the input surface normal $\hat{\mathbf{n}}'$ is computed using a learning-based UPS [3], which is the extended version of [2] used in the original PS-NeRF. For MonoSDF, we use a monocular normal and depth estimation

method [6] to obtain per-view normal and depth maps from a single image. We input the estimated normal maps as $\hat{\mathbf{n}}'$ and also incorporate a scale- and shift-invariant loss, proposed in MonoSDF, between the estimated depth and the rendered depth of the SDF. The weight of the depth loss is set to be the same as that of the normal loss.

Implementation details We implement the proposed method based on the official implementation of NeuS[1]. We train the model for $100K$ iterations with batch size 1024. The training takes about 10 hours on a NVIDIA A6000 GPU. We use Adam optimizer with default parameters. We use different learning rates for updating parameters of MLPs, θ and ϕ and the ambiguity matrix $\mathbf{X}$, set to 5×10^{-4} and 1×10^{-3}, respectively.

We set the weights of losses as $\lambda_{\text{mask}} = 0.1$ and $\lambda_{\text{reg}} = 0.1$. The weight of the normal loss λ_{normal} is initially set to 0.03 and linearly increased to 0.1 over the first $50K$ iterations. The coarse-to-fine optimization via positional encoding is also adapted from 0 to $50K$ iterations. To compute the confidence map, we store the change of loss, Eq. (9), every 500 iterations with downscaled resolution. The update step α in Eq. (9) is set to the learning rate for the ambiguity matrix. For the first $20K$ iterations, we do not use the confidence weighting to ensure stable computation of the variance. To avoid collapsed estimation, we fix the weight of the normal loss λ_{normal} for PS-NeRF and MonoSDF to 0.03.

Evaluation metrics For evaluation, we extract a mesh from the optimized SDF using the Marching cubes algorithm [24]. Invisible surfaces from any cameras are excluded using a rasterizer-based renderer [32], and we compute the Chamfer distance between the point clouds extracted from the estimated and the ground truth meshes. We additionally evaluate the angular error between the rendered normal maps of the optimized SDF and the ground truth.

4.1. Synthetic Experiments

We render two scenes for our evaluation, BLOBBY [10] and BUNNY[2], using a rendering software, Blender[3]. We use 24 viewpoints, rotating the camera at equal intervals around the object, and for each viewpoint, render 3 images under different directional lights, which move together with the camera. As shown in Fig. 4, we use a textureless surface for BLOBBY and a wood texture[4] for BUNNY.

As discussed in [3], specularity is useful for accurate estimation of UPS; hence, we include specularities in the input images for PS-NeRF and MonoSDF. Conversely, the proposed method and NeuS assume the Lambertian surface, and thus, we render them as diffuse surfaces.

We use four views located at 90-degree intervals and three light directions. Figure 7 and Table 1 present the estimated normal maps and estimation errors, respectively. Here, we visualize only one view per scene, and more complete results can be found in our supplementary material. Due to sparse viewpoints, NeuS fails to reconstruct correct surfaces. In contrast, MonoSDF, which uses the same inputs as NeuS, can accurately recover the rough shapes. However, MonoSDF tends to produce over-smoothed surfaces, as observed in the BUNNY scene. PS-NeRF provides a reasonable estimation, but due to the lower accuracy of surface normal estimation by UPS, the accuracy of its optimized shapes is not as high as that of the proposed method. In terms of both the Chamfer distances and mean angular errors, we observe the accurate estimation achieved by the proposed method.

4.2. Real-world Experiments

We use two datasets for real-world evaluation, DiLiGenT-MV [19] and our own dataset. We first describe the datasets and then introduce the experimental results.

DiLiGenT-MV: DiLiGenT-MV captures five objects from 20 viewpoints and uses fixed 96 light directions. Since the proposed method assumes the Lambertian reflectance, we here show the results for two scenes with relatively diffuse materials, BEAR and BUDDHA shown in Fig. 6. The remaining scenes are shown in our supplementary material.

Our dataset: For recording real-world data, we use a polarimetric camera (FLIR BFS-U3-51S5PC-C) for obtaining the diffuse-only and diffuse+specular observations for comparison methods. For just running our method, in practice, we can simply attach an off-the-shelf polarization filter to an ordinary camera, as shown in Fig. 1.

We capture three objects: RABBIT, CUBE, and FROG, shown in Fig. 6. RABBIT and FROG are made of pottery, while CUBE is made of plastic. We use our capturing setup, shown in Fig. 5, to capture images from 60 viewpoints with 3 light directions, rotating the target object. The ground truth is obtained by a laser scanner and aligned with the estimated meshes by NeuS, using all of the 60 views.

Results: Figures 7 and 8 visualize the estimated normal maps and meshes, respectively, and Table 1 represents the estimation errors. Similar to our synthetic experiments, NeuS presents larger errors for most of the scenes, and PS-NeRF faces challenges in recovering the normal map from limited observations. Although our dataset (RABBIT, CUBE, and FROG) is captured using cross-polarization to reduce specular reflections, BEAR and BUDDHA contain specu-

[1] NeuS implementation, https://github.com/Totoro97/NeuS/, last accessed on March 25, 2024.
[2] Stanford Bunny, https://graphics.stanford.edu/data/3Dscanrep/, last accessed on March 25, 2024.
[3] Blender 3.3, https://www.blender.org/, last accessed on March 25, 2024.
[4] Poly Haven, https://polyhaven.com/, last accessed on March 25, 2024.

Table 1. Comparison results with four views. PS-NeRF and ours use three light sources. "CD" denotes the Chamfer distance (↓) between the mesh extracted from the estimated SDF and the ground truth. "MAE" denotes mean angular errors (↓) in degrees across all available views. We show the MAE of rendered normal maps of the estimated SDF and estimated normal maps fed to the optimization. The estimated normal maps by the proposed method are disambiguated by the estimated ambiguity matrix $\mathbf{X}$. Bold font and underline are used to denote the best and second-best results, respectively.

	NeuS		MonoSDF			PS-NeRF			Ours		
	CD	MAE (SDF)	CD	MAE (SDF)	MAE (Est.)	CD	MAE (SDF)	MAE (Est.)	CD	MAE (SDF)	MAE (Est.)
BLOBBY	<u>24.3</u>	14.0	27.3	8.40	30.9	25.3	<u>7.05</u>	<u>16.0</u>	**8.83**	**2.28**	**7.01**
BUNNY	16.3	21.9	7.61	11.5	20.6	<u>5.38</u>	<u>8.53</u>	<u>17.1</u>	**1.79**	**5.38**	**12.4**
BEAR	10.2	10.4	<u>1.7</u>	<u>5.81</u>	<u>13.0</u>	6.32	7.02	18.1	**1.30**	**5.11**	**12.9**
BUDDHA	26.6	30.1	<u>12.2</u>	23.1	<u>26.2</u>	13.4	<u>22.5</u>	29.8	**2.18**	**14.5**	**19.3**
RABBIT	<u>2.17</u>	<u>13.1</u>	2.20	14.2	<u>26.2</u>	3.81	13.4	29.8	**1.65**	**9.67**	**15.2**
CUBE	95.2	34.6	<u>0.90</u>	<u>9.64</u>	<u>14.1</u>	9.50	14.7	32.6	**0.85**	**8.34**	**12.6**
FROG	40.1	33.7	20.2	23.6	26.2	<u>13.5</u>	<u>19.3</u>	<u>25.6</u>	**2.01**	**14.8**	**18.2**

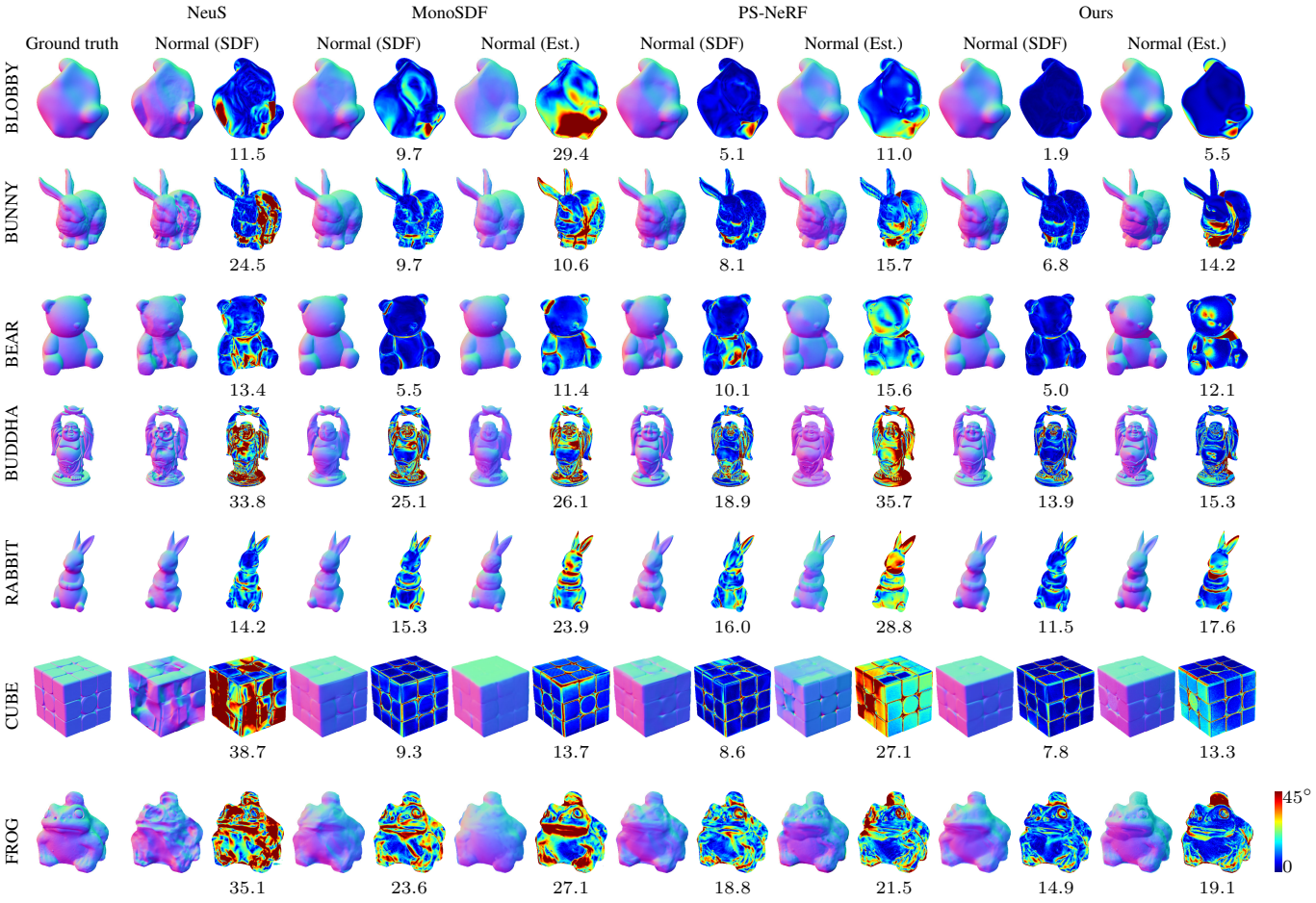

Figure 7. Evaluation of the normal maps. For each scene and method, we present the rendered normal map of the SDF, the estimated normal map fed to the optimization, and corresponding error maps side-by-side. The numbers under the error maps represent mean angular errors in degrees. The estimated normal maps by the proposed method are disambiguated by the estimated ambiguity matrix $\mathbf{X}$.

lar reflections. Nevertheless, the proposed method robustly estimates accurate shapes.

In scenes such as BEAR and CUBE, the monocular normal estimation in MonoSDF performs well, leading to accurate estimations. However, for example, in the CUBE scene, the detailed shape is lacking. In contrast, the proposed method can recover detailed shapes in the CUBE scene. BUDDHA and FROG present challenges for all methods; however, the proposed method achieves globally accurate shape recovery.

Table 2. Comparison of the proposed method with and without HO-GSVD. "MAE (UPS)" and "MAE (SDF)" denote the mean angular errors (↓) in degrees for the disambiguated normal maps and rendered normal maps of SDF, respectively. "Proj. mat." represents the error between the estimated ambiguity matrix and the ground truth measured by the Frobenius norm (↓). Bold font indicates better results.

	4 views			8 views			20 views		
	MAE (UPS)	Proj. mat.	MAE (SDF)	MAE (UPS)	Proj. mat.	MAE (SDF)	MAE (UPS)	Proj. mat.	MAE (SDF)
Ours	**0.13**	**0.0023**	**1.78**	**0.085**	**0.0010**	**1.32**	**0.054**	**0.00055**	**1.29**
w/o HO-GSVD	0.62	0.0050	3.00	0.59	0.0037	2.01	0.421	0.0026	1.77

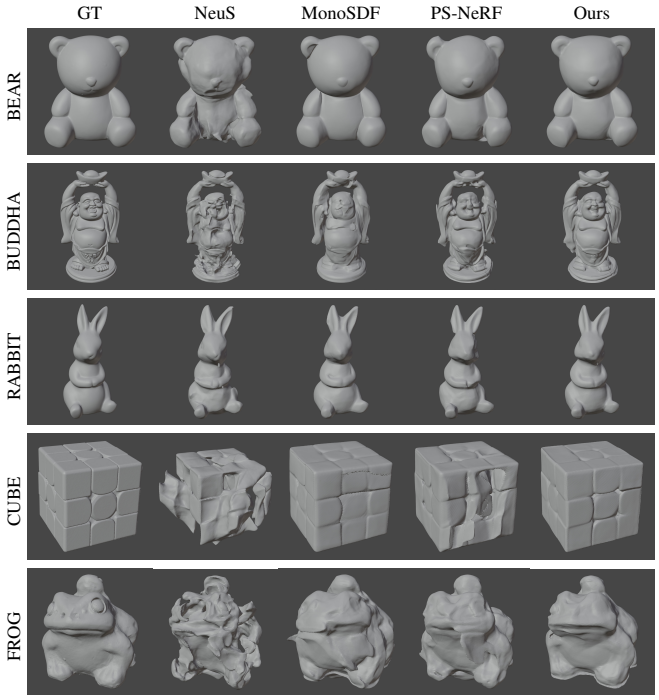

Figure 8. Estimated meshes for the real-world scenes.

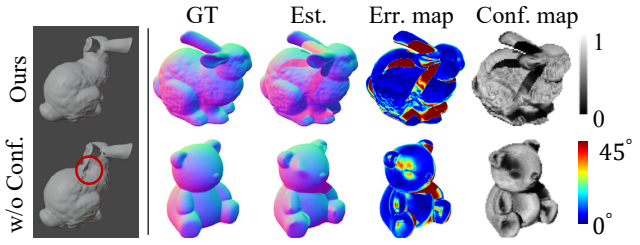

Figure 9. Results of confidence estimation. The left-hand side shows the estimated meshes for the BUNNY scene, both ours and ours without confidence estimation ("w/o Conf."). On the right-hand side, for each row, the first two columns present the normal maps of the ground truth and the one estimated, projected by the estimated ambiguity matrix. The last two columns show the error maps ("Err. map") of the estimated normal and confidence maps ("Conf. map").

5. Discussion

This paper presents a practical and easy-to-implement 3D reconstruction method, MVCPS. Our constrained setting allows us to decompose the observations into per-view surface normal maps and shared light directions w.r.t. the camera by HO-GSVD, reducing the per-view ambiguities in UPS to a single and global linear ambiguity. We demonstrate that, by integrating the decomposed normal maps into neural surface reconstruction, the proposed method can jointly estimate accurate 3D shapes and the ambiguity matrix. We compare the proposed method with the state-of-the-art methods and show the proposed method's effectiveness in the challenging setting of sparse views and lights.

One of the limitations of our method is the Lambertian assumption, which is rarely met in the real world. Dealing with non-Lambertian objects under sparse viewpoints and light sources is one of our future venues.

Ablation study on HO-GSVD To assess the impact of HO-GSVD compared to SVD-based UPS, we evaluate the accuracy of disambiguation in the proposed optimization stage. We assume that the factorization perfectly works and use the normal maps of the ground truth as input. We use the BUNNY scene with 4, 8, and 20 views. Table 2 compares the disambiguated accuracy of the surface normal using a shared linear ambiguity across all views (HO-GSVD), with those using per-view linear ambiguities (SVD). Since we can only disambiguate the surface normals of the views used in the optimization when using view-independent ambiguities, we compare the mean angular errors for those specific views in this experiment. This result demonstrates the consistent advantage of HO-GSVD for more accurate ambiguity resolution. We further investigate the effectiveness of HO-GSVD in improving factorization accuracy, with details provided in our supplementary material.

Ablation study on confidence estimation Figure 9 shows the results of the proposed method with and without confidence estimation. We also visualize the disambiguated normal maps, the corresponding error maps, and the estimated confidence maps. As seen in the error maps and the confidence maps, pixels with higher errors tend to have lower confidence, as expected. In the BEAR scene, both shadowed and specular pixels exhibit lower confidence, which contributes to robust estimation. Observing the estimated mesh without confidence estimation reveals that the shape is heavily affected by the shadow on the ear.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP22K17910 and JP23H05491.

References

[1] P.N. Belhumeur, D.J. Kriegman, and A.L. Yuille. The bas-relief ambiguity. In *Computer Vision and Pattern Recognition (CVPR)*, 1997. 2

[2] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K. Wong. SDPS-Net: Self-calibrating deep photometric stereo networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[3] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K. Wong, and Yasuyuki Matsushita. What is learned in deep uncalibrated photometric stereo? In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6

[4] Ondrej Drbohlav and Mike Chaniler. Can two specular pixels calibrate photometric stereo? In *International Conference on Computer Vision (ICCV)*, 2005. 2

[5] Ondřej Drbohlav and Radim Šára. Specularities reduce ambiguity of uncalibrated photometric stereo. In *European Conference on Computer Vision (ECCV)*, 2002. 2

[6] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 6

[7] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376, 2009. 2

[8] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning (ICML)*, 2020. 4

[9] Hideki Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of the Optical Society of America (JOSA)*, 11(11):3079–3089, 1994. 2, 3

[10] Micah K. Johnson and Edward H. Adelson. Shape estimation in natural illumination. In *Computer Vision and Pattern Recognition (CVPR)*, 2011. 6

[11] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncalibrated neural inverse rendering for photometric stereo of general surfaces. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[13] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool. Neural radiance fields approach to deep multi-view photometric stereo. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022. 2

[14] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Multi-view photometric stereo revisited. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. 1, 2

[15] Idris Kempf, Paul J. Goulart, and Stephen R. Duncan. A higher-order generalized singular value decomposition for rank-deficient matrices. *SIAM Journal on Matrix Analysis and Applications*, 44(3):1047–1072, 2023. 3

[16] David J. Kriegman and Peter N. Belhumeur. What shadows reveal about object structure. *Journal of the Optical Society of America (JOSA)*, 18(8):1804–1813, 2001. 2

[17] JH Lambert. Photometria. *Augustae Vindelicorum*, 1760. 2

[18] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[19] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020. 1, 2, 4, 6

[20] Zongrui Li, Qian Zheng, Boxin Shi, Gang Pan, and Xudong Jiang. Dani-net: Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[21] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[22] F. Logothetis, R. Mecca, and R. Cipolla. A differential volumetric approach to multi-view photometric stereo. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[23] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. SparseNeuS: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[24] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *SIGGRAPH*, New York, NY, USA, 1987. Association for Computing Machinery. 6

[25] Lilika Makabe, Heng Guo, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Near-light photometric stereo with symmetric lights. In *IEEE International Conference on Computational Photography (ICCP)*, 2023. 1

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 4

[27] Kazuma Minami, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Symmetric-light photometric stereo. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022. 1

[28] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[29] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Multiview photometric stereo using planar mesh parameterization. In *International Conference on Computer Vision (ICCV)*, 2013. 1, 2

[30] Sri Ponnapalli, Michael Saunders, Charles Loan, and Orly Alter. A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms. *PloS one*, 6:e28072, 2011. 3

[31] Yvain Quéau, Bastien Durix, Tao Wu, Daniel Cremers, François Lauze, and Jean-Denis Durou. LED-based photometric stereo: Modeling, calibration and numerical solution. *Journal of Mathematical Imaging and Vision*, 60(3):313–340, 2018. 1

[32] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501*, 2020. 6

[33] Hiroaki Santo, Michael Waechter, and Yasuyuki Matsushita. Deep near-light photometric stereo for spatially varying reflectances. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[34] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo networks for determining surface normal and reflectances. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(1): 114–128, 2022.

[35] Boxin Shi, Zhipeng Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(2):271–284, 2019. 2

[36] William M. Silver. Determining shape and reflectance using multiple images. Master's thesis, Massachusetts Institute of Technology, 1980. 2

[37] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 4, 5

[38] Robert J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980. 2

[39] Haoyu Wu, Alexandros Graikos, and Dimitris Samaras. S-VolSDF: Sparse multi-view stereo regularization of neural implicit surfaces. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[40] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. PS-NeRF: Neural inverse rendering for multi-view photometric stereo. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 5

[41] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[42] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Neural Information Processing Systems (NeurIPS)*, 2021. 2

[43] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. In *Neural Information Processing Systems (NeurIPS)*, 2022. 2, 5

[44] Dongxu Zhao, Daniel Lichy, Pierre-Nicolas Perrin, Jan-Michael Frahm, and Soumyadip Sengupta. MVPSNet: Fast generalizable multi-view photometric stereo. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[45] Zhenglong Zhou and Ping Tan. Ring-light photometric stereo. In *European Conference on Computer Vision (ECCV)*, 2010. 2