Multi-View Azimuth Stereo via Tangent Space Consistency

Xu Cao Hiroaki Santo Fumio Okura Yasuyuki Matsushita
Osaka University
{cao.xu, santo.hiroaki, okura, yasumat}@ist.osaka-u.ac.jp

Source code: https://github.com/xucao-42/mvas

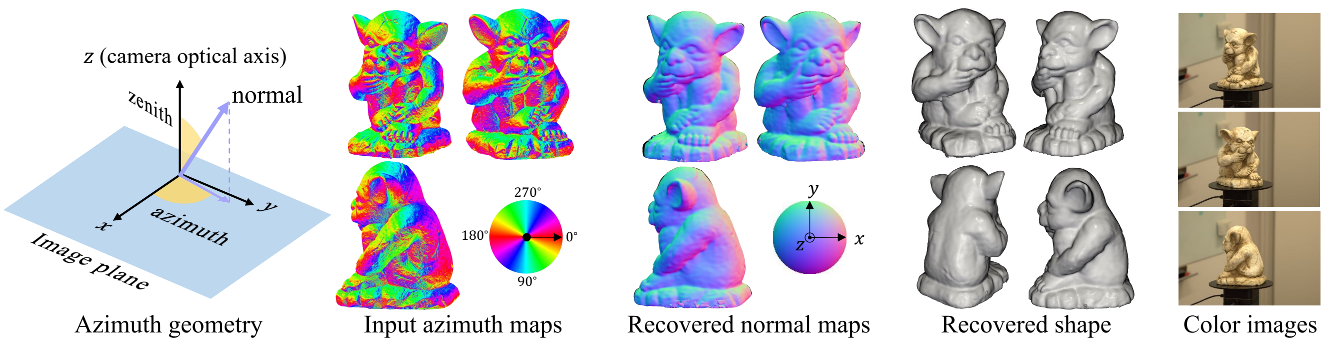


Figure 1. 3D reconstruction from calibrated multi-view azimuth maps (3 out of 31 are shown). An azimuth angle indicates the surface normal's orientation in the image plane, and an azimuth map records the azimuth angles across the entire surface. We show that azimuth maps can be effectively used for shape and normal recovery. Color images are for reference only and are not used in shape optimization.

Abstract

We present a method for 3D reconstruction only using calibrated multi-view surface azimuth maps. Our method, multi-view azimuth stereo, is effective for textureless or specular surfaces, which are difficult for conventional multi-view stereo methods. We introduce the concept of tangent space consistency: Multi-view azimuth observations of a surface point should be lifted to the same tangent space. Leveraging this consistency, we recover the shape by optimizing a neural implicit surface representation. Our method harnesses the robust azimuth estimation capabilities of photometric stereo methods or polarization imaging while bypassing potentially complex zenith angle estimation. Experiments using azimuth maps from various sources validate the accurate shape recovery with our method, even without zenith angles.

1. Introduction

Recovering 3D shapes of real-world scenes is a fundamental problem in computer vision, and multi-view stereo (MVS) has emerged as a mature geometric method for reconstructing dense scene points. Using 2D images taken from different viewpoints, MVS finds dense correspondences between images based on the photo-consistency assumption, that a scene point's brightness should appear similar across different viewpoints [13, 37–39]. However, MVS struggles with textureless or specular surfaces, as the lack of texture leads to ambiguities in establishing correspondences, and the presence of specular reflections violates the photo-consistency assumption [11].

Photometric stereo (PS) offers an alternative approach for dealing with textureless and specular surfaces [32]. By estimating single-view surface normals using varying lighting conditions [40], PS enables high-fidelity 2.5D surface reconstruction [26]. However, extending PS to a multi-view setup, known as multi-view photometric stereo (MVPS) [15], significantly increases image acquisition costs, as it requires multi-view and multi-light images under highly controlled lighting conditions [21].

To mitigate image acquisition costs, simpler lighting setups such as circularly or symmetrically placed lights have been explored [2, 3, 23, 44]. With these lighting setups, estimating the surface normal's azimuth (the angle in the image plane) becomes considerably easier than estimating the zenith (the angle from the camera optical axis) [2, 3, 23]. The ease of azimuth estimation also appears in polarization imaging [29]. While azimuth can be determined up to a π-ambiguity using only polarization data, zenith estimation requires more complex steps [24, 34, 36].

In this paper, we introduce Multi-View Azimuth Stereo (MVAS), a method that effectively uses calibrated multi-view azimuth maps for shape recovery (Fig. 1). MVAS is particularly advantageous when working with accurate azimuth acquisition techniques. With circular-light photometric stereo [3], MVAS has the potential to be applied to surfaces with arbitrary isotropic materials. With polarization imaging [7], MVAS allows a passive image acquisition as simple as MVS while being more effective for textureless or specular surfaces.

The key insight enabling MVAS is the concept of Tangent Space Consistency (TSC) for multi-view azimuth angles. We find that the azimuth can be transformed into a tangent using camera orientation. Therefore, multi-view azimuth observations of the same surface point should be lifted to the same tangent space (Fig. 2). TSC helps determine if a 3D point lies on the surface, similar to photo-consistency for finding image correspondences. Moreover, TSC can directly determine the surface normal as the vector orthogonal to the tangent space, enabling high-fidelity reconstruction comparable to MVPS methods. Notably, TSC is invariant to the π-ambiguity of the azimuth angle, making MVAS well-suited for polarization imaging.

With TSC, we reconstruct the surface implicitly represented as a neural signed distance function (SDF), by constraining the surface normals (*i.e.*, the gradients of the SDF). Experimental results show that MVAS achieves comparable reconstruction performance to MVPS methods [18, 28, 41], even in the absence of zenith information. Further, MVAS outperforms MVS methods [31] in textureless or specular surfaces using azimuth maps from symmetric-light photometric stereo [23] or a snapshot polarization camera [7].

In summary, this paper's key contributions are:

- Multi-View Azimuth Stereo (MVAS), which enables accurate shape reconstruction even for textureless and specular surfaces;
- Tangent Space Consistency (TSC), which establishes the correspondence between multi-view azimuth observations, thereby facilitating the effective use of azimuth data in 3D reconstruction; and
- A comprehensive analysis of TSC, including its necessary conditions, degenerate scenarios, and the application to optimizing neural implicit representations.

2. Related Tasks and Concept

This section discusses the relation of MVAS to multi-view photometric stereo (MVPS) and shape-from-polarization (SfP), and compares TSC to photo-consistency.

MVPS versus MVAS MVPS aims for high-fidelity shape and reflectance recovery using images from different angles and under different lighting conditions [15, 22]. These "multi-light" images can be used for estimating and fusing multi-view normal maps [4, 18], for refining coarse meshes initialized by MVS [28], or for jointly estimating the shape and materials in an inverse-rendering manner [41].

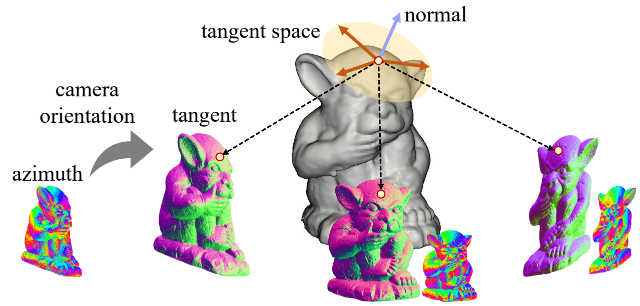


Figure 2. Tangent space consistency. The azimuth can be converted to a tangent by camera orientation. The tangents in different views, but projected from the same surface point, should lie in the same tangent space and can directly determine the surface normal.

Compared to MVPS, MVAS has the potential to be applied to (1) surfaces of a broader range of materials and/or (2) in uncontrolled scenarios, benefiting from azimuth inputs. First, azimuth estimation is valid for arbitrary isotropic materials using an uncalibrated circular moving light [3], while MVPS methods require specific surface reflectance modeling (*e.g.*, Lambertian [4] or the microfacet model [41]) or prior learning [18]. Second, MVAS allows passive image capture with polarization imaging, while MVPS has to actively illuminate the scene, limiting MVPS's application in highly controlled environments.

SfP versus MVAS SfP recovers surfaces using polarization imaging [1]. For dielectric surfaces, the measured angle of polarization (AoP) aligns with the surface normal's azimuth component, up to a π ambiguity. SfP studies determine surface normals by resolving this π-ambiguity and estimating the zenith component [8–10, 16, 17, 29, 34, 35, 45]. Some studies use polarization data to refine coarse shapes initialized by multi-view reconstruction methods [6, 43], but the geometric relation between multi-view azimuth angles are not considered.

With TSC and MVAS, both the π-ambiguity and zenith estimation can be bypassed. Our method relies on TSC, not requiring MVS methods to initialize shapes.

Photo-consistency versus tangent space consistency Photo-consistency is a key assumption in MVS for establishing correspondence between multi-view images. This assumption states that a scene point appears similar across different views and struggles with specular surfaces [12].

In contrast, TSC is derived from geometric principles and strictly holds for multi-view azimuth angles. Further, TSC can determine the surface normal, providing more information than photo-consistency. However, TSC requires at least three cameras with non-parallel optical axes and can degrade to photo-consistency under certain camera configurations. Similar to photo-consistency's challenges with textureless surfaces, TSC might struggle to establish correspondences for planar surfaces. Details are in Sec. 3.2.

3. Proposed Method

We aim to recover the shape from calibrated and masked azimuth maps. Let Ω_i represent the i-th image pixel domain. For each view $i \in \{1, 2, ..., C\}$, we assume the following are available:

- a surface azimuth map $\phi_i : \Omega_i \to [0, 2\pi]$,
- a binary mask indicating whether a pixel is inside the shape silhouette $O_i : \Omega_i \to \{0, 1\}$, and
- the projection from the world coordinates to the image pixel coordinates $\Pi_i : \mathbb{R}^3 \to \Omega_i$, consisting of the extrinsic rigid-body transformation $\mathbf{P}_i = [\mathbf{R}_i \mid \mathbf{t}_i] \in SE(3)$ and intrinsic perspective camera projection $\mathbf{K}_i$.

We describe the proposed method in three sections. First, we detail the transformation from an azimuth angle to a projected tangent vector (Sec. 3.1). Next, we discuss multi-view tangent space consistency for surface points, including its four degenerate scenarios and π-invariance (Sec. 3.2). Lastly, we present the surface reconstruction by optimizing a neural implicit representation based on the tangent space consistency loss (Sec. 3.3).

3.1. The projected tangent vector

This section will show how to convert an azimuth angle to a tangent vector of the surface point, given the world-to-camera rotation. We will only consider single-view observations and ignore the view index in this section.

In the world coordinates, consider a unit normal vector $\mathbf{n}(\mathbf{x}) \in \mathcal{S}^2 \subset \mathbb{R}^3$ of a surface point $\mathbf{x} \in \mathbb{R}^3$. Suppose a rigid-body transformation $[\mathbf{R} \mid \mathbf{t}]$ transforms the surface from the world coordinates to the camera coordinates. The direction of the normal vector in the camera coordinates $\mathbf{n}^c$ is rotated accordingly as

$$\mathbf{R}\mathbf{n} = \mathbf{n}^c. \tag{1}$$

In the camera coordinates, we can parameterize the unit normal vector by its azimuth angle $\phi \in [0, 2\pi]$ and zenith angle $\theta \in \left[0, \frac{\pi}{2}\right)$ as

$$\mathbf{n}^c = \begin{bmatrix} n_x^c \\ n_y^c \\ n_z^c \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix}. \tag{2}$$

From Eq. (2), we can derive the relation between n_x^c and n_y^c in terms of only the azimuth angle as

$$n_x^c \sin\phi = n_y^c \cos\phi. \tag{3}$$

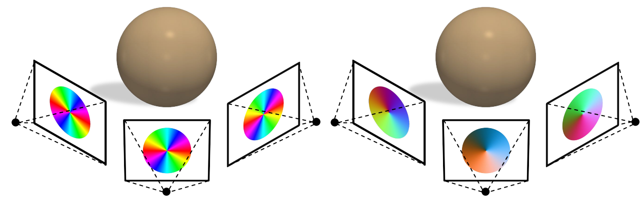


Figure 3. The azimuth angle observations **(left)** are lifted to tangent vectors **(right)** by world-to-camera rotations. The tangent vectors are coded by 8-bit RGB colors using $255(\mathbf{t} + \mathbf{1})/2$.

Denoting the rotation matrix as

$$\mathbf{R} = \begin{bmatrix} -\mathbf{r}_1^\top- \\ -\mathbf{r}_2^\top- \\ -\mathbf{r}_3^\top- \end{bmatrix} \in SO(3), \tag{4}$$

and putting Eqs. (1) to (4) together, we obtain

$$\mathbf{r}_1^\top \mathbf{n} \sin\phi = \mathbf{r}_2^\top \mathbf{n} \cos\phi. \tag{5}$$

Rearranging Eq. (5) yields

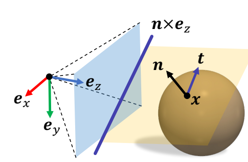
$$\mathbf{n}^\top \underbrace{(\mathbf{r}_1 \sin\phi - \mathbf{r}_2 \cos\phi)}_{\mathbf{t}(\phi)} = 0. \tag{6}$$

We call $\mathbf{t}(\phi)$ the *projected tangent vector*, as it is computed from the projected azimuth angle and perpendicular to the surface normal. As shown in Figure 3, the transformation from azimuth maps to tangent maps reveals that projected tangent vectors encode camera orientation information, providing useful hints for multi-view reconstruction.

Properties The projected tangent vector is the unit vector parallel to the intersection of the tangent and image spaces. Based on Eq. (6),

$$\begin{aligned} \mathbf{t}^\top\mathbf{t} &= \mathbf{r}_1^\top\mathbf{r}_1 \sin^2\phi + \mathbf{r}_2^\top\mathbf{r}_2 \cos^2\phi - 2\mathbf{r}_1^\top\mathbf{r}_2 \cos\phi\sin\phi \\ &= \sin^2\phi + \cos^2\phi = 1, \end{aligned} \tag{7}$$

since $\mathbf{r}_1$ and $\mathbf{r}_2$ are orthonormal vectors.



The inset illustrates the second property. Let $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$ be the unit direction vector of the x-, y-, and z-axis of the camera coordinates in the world coordinates. Then $\mathbf{r}_1 = \mathbf{e}_x$ and $\mathbf{r}_2 = \mathbf{e}_y$, which follows that $\mathbf{t}(\phi)$ is a linear combination of camera's x- and y-axes and thus parallel to the image plane. We can compute the intersection direction of two planes by taking the cross-product of their normals, namely, the surface normal and the principle axis. Hence,

$$\mathbf{t} \parallel \mathbf{n} \times \mathbf{e}_z. \tag{8}$$

The two properties are helpful in analyzing the tangent space consistency, as described next.

3.2. Multi-view tangent space consistency

This section discusses the consistency between multi-view azimuth observations in the tangent space of a surface point. In addition, four degenerate scenarios and π-invariance will be discussed. We assume the surface point under consideration is visible to all cameras in this section.

Denote the projected tangent vector of a surface point in i-th view as $\mathbf{t}_i(\mathbf{x}) = \mathbf{t}(\phi_i(\Pi_i(\mathbf{x})))$. By Eq. (6), a surface point $\mathbf{x}$, its normal direction $\mathbf{n}$, and its multi-view projected tangent vectors $\mathbf{t}_i$ should satisfy:

$$\mathbf{n}(\mathbf{x})^\top \mathbf{t}_i(\mathbf{x}) = 0 \quad \forall i. \tag{9}$$

Let $\mathbf{T}(\mathbf{x}) = [\mathbf{t}_1(\mathbf{x}), \mathbf{t}_2(\mathbf{x}), ..., \mathbf{t}_C(\mathbf{x})]^\top \in \mathbb{R}^{C\times 3}$ be the matrix formed by stacking projected tangent vectors of all C views. Then Eq. (9) reads

$$\mathbf{T}(\mathbf{x})\mathbf{n}(\mathbf{x}) = \mathbf{0}. \tag{10}$$

Equation (10) can only be satisfied if the rank of $\mathbf{T}(\mathbf{x})$ is either 1 or 2. The rank cannot be 0 as projected tangent vectors are unit length. The case $\text{rank}(\mathbf{T}(\mathbf{x})) = 3$ cannot satisfy Eq. (10) as surface normals are non-zero vectors.

We refer to the case where the rank of $\mathbf{T}(\mathbf{x})$ is 2 as *tangent space consistency* (TSC). In this case, multi-view projected tangent vectors from a surface point span its tangent space, and the surface normal is determined up to a sign ambiguity. On the other hand, when $\text{rank}(\mathbf{T}(\mathbf{x})) = 1$, the projected tangent vectors can only span a tangent line and constrain the surface normal on the plane orthogonal to the tangent line. This can occur when camera optical axes are parallel, as explained later.

TSC can help distinguish non-surface points (wrong correspondences) from surface points (possibly correct correspondences) and determine the surface normals, as shown in Fig. 4. For wrong correspondences, their projected tangent vectors are expected to have a rank of 3 and span the entire 3D space. On the other hand, for surface points, their projected tangent vectors span the tangent space, i.e., $\text{rank}(\mathbf{T}(\mathbf{x})) = 2$. In addition, TSC requires the surface normal to be in the null space of $\mathbf{T}(\mathbf{x})$, i.e., perpendicular to the tangent space spanned by projected tangent vectors. This makes TSC more informative than photo-consistency since photo-consistency cannot directly determine the surface normal.

To effectively distinguish surface/non-surface points using TSC, a non-planar surface must be observed by at least three cameras with non-parallel optical axes. These requirements indicate four degeneration scenarios, as shown in Fig. 5 and discussed below. Table 1 summarizes the variations of $\text{rank}(\mathbf{T}(\mathbf{x}))$ in these scenarios.

Number of viewpoints For TSC to be effective, the rank of $\mathbf{T}(\mathbf{x})$ is expected to be 3 for non-surface points. However, when only two views are available, the rank of $\mathbf{T}(\mathbf{x})$ is impossible to achieve 3 since $\mathbf{T}(\mathbf{x}) \in \mathbb{R}^{2\times 3}$. In this case, $\text{rank}(\mathbf{T}(\mathbf{x})) \leq 2$ is satisfied for arbitrary correspondence. Consequently, TSC cannot distinguish surface points from non-surface points in the two-view case.

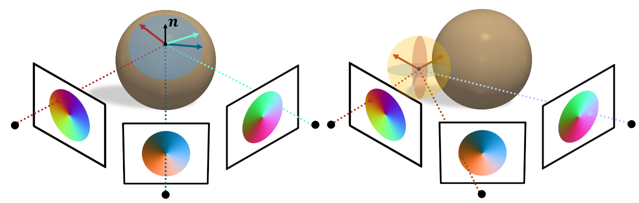


Figure 4. **(Left)** Multi-view projected tangent vectors from a surface point span its tangent space and determine the surface normal. **(Right)** Conversely, multi-view projected tangent vectors from a non-surface point (*i.e.*, wrong correspondence) are expected to span the 3D space.

Table 1. Rank of $\mathbf{T}(\mathbf{x})$ in four degenerate cases of TSC.

Scenarios	Non-surface points	surface points	surface normal
Two-view	2	2	×
Co-linear optical axes	2	1	×
Co-planar optical axes	2	1	△
Planar surface	2	2	✓
TSC	3	2	✓

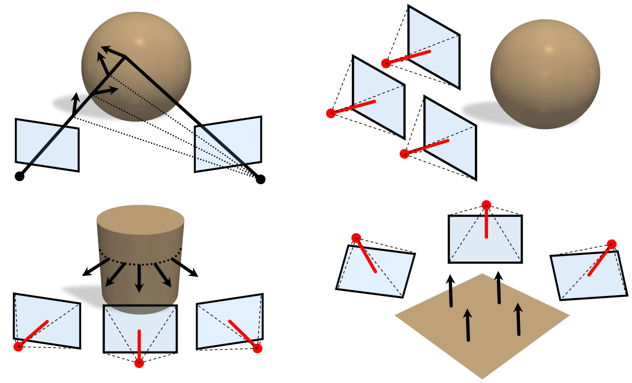


Figure 5. Degeneration scenarios where TSC cannot distinguish good correspondence from bad ones. **(Top Left)** Two-view observations, **(Top Right)** frontal parallel cameras with parallel optical axes (red pins), **(Bottom Left)** cameras with coplanar optical axes observe coplanar surface normals, and **(Bottom Right)** planar surface regions.

Camera setups TSC requires the projected tangent vectors of a surface point can span the tangent space but not a tangent line. This requirement breaks down when projected tangent vectors are observed from (1) frontal parallel cameras, or (2) cameras with coplanar optical axes.

Frontal parallel cameras have parallel optical axes. By Eq. (8), multi-view projected tangent vectors of a surface point also become parallel. This reduces the rank of $\mathbf{T}(\mathbf{x})$ to 1, and TSC degrades to photo-consistency since all cameras should observe the same tangent vector for a surface point.

A more special case is when cameras with coplanar optical axes observe coplanar surface normals, such as a rotating camera observing a cylinder. In this case, the cross product of the coplanar normal and optical axis vectors yields co-linear projected tangent vectors. As such, the rank of $\mathbf{T}(\mathbf{x})$ is 1 for surface points, and TSC again degrades to photo-consistency. However, this degradation does not occur for non-coplanar surface normals, meaning TSC can still be effective for general surfaces.

Surface types TSC breaks down for a planar surface. At any location on the planar surface, $\mathbf{n}(\mathbf{x})$ is the same and $\mathrm{rank}(\mathbf{T}(\mathbf{x}))$ is identically 2 for arbitrary correspondence. However, the normal direction of this plane can still be correctly determined in the case $\mathrm{rank}(\mathbf{T}(\mathbf{x})) = 2$, *i.e.*, at least three non-frontal parallel views. The planar surface can be seen as the counterpart to the textureless region for photo-consistency. However, unlike photo-consistency, TSC can still determine the surface normal[1].

π-invariance TSC remains effective when the azimuth angle is changed by π. By Eq. (6), the sign of the projected tangent vector will be reversed:

$$\mathbf{t}(\phi+\pi) = -\mathbf{r}_1 \sin\phi + \mathbf{r}_2 \cos\phi = -\mathbf{t}(\phi). \tag{11}$$

Intuitively, reversing the direction of a tangent vector still places it in the same tangent space, as $\mathbf{n}^\top(-\mathbf{t}) = 0$ when $\mathbf{n}^\top \mathbf{t} = 0$. Mathematically, reversing the signs of arbitrary rows in $\mathbf{T}(\mathbf{x})$ does not affect the rank of $\mathbf{T}(\mathbf{x})$. This π-invariance can be particularly useful for polarization imaging, as they can only measure azimuth angles up to a π ambiguity.

3.3. Multi-view azimuth stereo

We propose the following TSC-based functional for multi-view geometry reconstruction:

$$\mathcal{J} = \oiint_{\mathcal{M}} \frac{\sum_{i=1}^{C} \Phi_i(\mathbf{x}) \left(\mathbf{n}(\mathbf{x})^\top \mathbf{t}_i(\mathbf{x})\right)^2}{\sum_{i=1}^{C} \Phi_i(\mathbf{x})} \, d\mathcal{M}. \tag{12}$$

Here, $\mathcal{M}$ is the surface embedded in the 3D space, and $d\mathcal{M}$ is the infinitesimal area on the surface. $\Phi_i(\mathbf{x})$ is a binary function indicating the visibility of the point $\mathbf{x}$ from the i-th viewpoint:

$$\Phi_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is visible to } i\text{-th camera} \\ 0 & \text{otherwise} \end{cases} . \tag{13}$$

We can simplify Eq. (12) as follows:

$$\mathcal{J} = \oiint_{\mathcal{M}} \mathbf{n}^\top \tilde{\mathbf{T}} \mathbf{n} \, d\mathcal{M} \quad \text{with} \quad \tilde{\mathbf{T}} = \frac{\sum_{i=1}^{C} \Phi_i \mathbf{t}_i \mathbf{t}_i^\top}{\sum_{i=1}^{C} \Phi_i}, \tag{14}$$

[1]A similar phenomenon exists in Helmholtz stereopsis [46], where wrong correspondence might still result in the correct normal estimation.

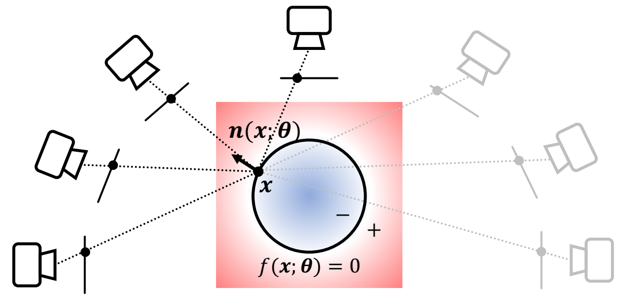


Figure 6. To optimize the neural SDF, we project the surface point onto all views and enforce the surface normal to be perpendicular to all visible projected tangent vectors.

where we omit the dependence on the surface point $\mathbf{x}$ for clarity. As discussed in Sec. 3.2, accurate surface points and normals are both necessary to minimize the functional.

We represent the surface implicitly using a signed distance function (SDF) and optimize the SDF based on the framework of implicit differentiable renderer (IDR) [42]. We parameterize the SDF by a multi-layer perceptron (MLP) as $f(\mathbf{x};\boldsymbol{\theta}) : \mathbb{R}^3 \times \mathbb{R}^d \to \mathbb{R}$, where $\mathbf{x} \in \mathbb{R}^3$ is the 3D point coordinate, and $\boldsymbol{\theta} \in \mathbb{R}^d$ are MLP parameters. The surface $\mathcal{M}$ is implicitly represented as the zero-level set of the SDF

$$\mathcal{M}(\boldsymbol{\theta}) = \{\mathbf{x} \mid f(\mathbf{x};\boldsymbol{\theta}) = 0\}, \tag{15}$$

which varies depending on the MLP parameters.

To optimize the MLP, we use a loss function that consists of the tangent space consistency loss, the silhouette loss, and the Eikonal regularization:

$$\mathcal{L} = \mathcal{L}_{\text{TSC}} + \lambda_1 \mathcal{L}_{\text{silhouette}} + \lambda_2 \mathcal{L}_{\text{Eikonal}}. \tag{16}$$

In each batch of the optimization, we randomly sample a set of P pixels from all views, cast camera rays from these pixels into the scene, and find the first ray-surface intersections. We evaluate the TSC loss for pixels with ray-surface intersections located inside the silhouette, denoted as $\mathbf{X}$. We evaluate the silhouette loss for pixels that do not have ray-surface intersections or are located outside the silhouette, denoted as $\tilde{\mathbf{X}}$.

Tangent space consistency loss Based on Eq. (14), we define the TSC loss as

$$\mathcal{L}_{\text{TSC}} = \frac{1}{P} \sum_{\mathbf{x}\in\mathbf{X}} \mathbf{n}(\mathbf{x};\boldsymbol{\theta})^\top \tilde{\mathbf{T}}(\mathbf{x}) \mathbf{n}(\mathbf{x};\boldsymbol{\theta}). \tag{17}$$

To evaluate the TSC loss, we need to evaluate the surface normal and construct the matrix $\tilde{\mathbf{T}}$. According to the property of SDF [27], the surface normal direction is the gradient evaluated at a zero-level set point:

$$\mathbf{n}(\mathbf{x};\boldsymbol{\theta}) = \nabla_{\mathbf{x}} f(\mathbf{x};\boldsymbol{\theta}). \tag{18}$$

Here, the surface normal can still be represented analytically as the MLP parameters [14, 33]. Therefore, the gradient of the loss functions can be backpropagated to MLP parameters via surface normals.

We then compute $\mathbf{T}(\mathbf{x})$ for the point $\mathbf{x}$ from all visible views. First, we project the surface points onto all views and check their visibility in each, as shown in Fig. 6. To determine the visibility, we march the surface points toward the camera center and check whether there is a negative distance on the ray; see the supplementary material for more details. Then in visible views, we compute the projected tangent vectors from input azimuth maps.

Silhouette loss Following IDR [42], we use the input masks to constrain the visual hull of the shape[2]. We find the minimal distance on the rays for pixels that do not have ray-surface intersections, denoted as f^*. The silhouette loss is then

$$\mathcal{L}_{\text{silhouette}} = \frac{1}{\alpha P} \sum_{\mathbf{x} \in \tilde{\mathbf{X}}} \Psi\big(O(\Pi(\mathbf{x})), \sigma(\alpha f^*)\big), \tag{19}$$

where Ψ is the cross entropy function, and $\sigma(\cdot)$ is a sigmoid function with α controlling its sharpness.

Eikonal regularization Following IGR [14], we use the Eikonal loss to regularize the gradient of SDF such that the gradient norm is close to 1 everywhere [27]:

$$\mathcal{L}_{\text{Eikonal}} = \mathbb{E}_{\mathbf{x}} \left((\|\mathbf{n}\|_2 - 1)^2 \right). \tag{20}$$

To apply Eikonal regularization, we randomly sample points within the object bounding box and compute the mean squared deviation from 1-norm.

None of the three loss functions explicitly constrain the surface points. It is the TSC loss that implicitly encourages good correspondence.

4. Experiments

We evaluate MVAS in three experiments: comparing with MVPS methods quantitatively for surface and normal reconstruction in Sec. 4.1, applying MVAS to a photometric stereo method which struggles with zenith estimation in Sec. 4.2, and using MVAS with passive polarization imaging in Sec. 4.3. Implementation details are in the supplementary material.

4.1. MVAS versus MVPS

Baselines We assess MVAS against multiple MVPS methods using the DiLiGenT-MV benchmark [21]. The MVPS methods include the coarse mesh refinement method R-MVPS [28], the benchmark method B-MVPS [21], the depth-normal fusion-based method UA-MVPS [18], and the neural inverse rendering method PS-NeRF [41]. DiLiGenT-MV [21] captures 20 views under 96 different lights for five objects. We use 15-view azimuth maps for optimization and leave out 5 views for testing, following PS-NeRF [41]. The azimuth maps are computed from the normal maps estimated by the self-calibrated photometric stereo method SDPS [5].

Evaluation metrics We use Chamfer distance (CD) and F-score for geometry accuracy [18, 19], and mean angular error (MAE) for normal accuracy [41]. For CD and F-score, we only consider visible points by casting rays for all pixels and finding the first ray-mesh intersections[3].

Results and discussions Table 2 reports the geometry accuracy of the recovered DiLiGenT-MV surfaces. B-MVPS [21] achieves the best scores in 4 objects due to the usage of calibrated light information. UA-MVPS [18] distorts the surface reconstruction by not considering the multi-view consistency. MVAS outperforms PS-NeRF [41] in 3 objects without modeling the rendering process.

Figure 7 visually compares recovered "Buddha" and "Reading" objects. Despite not having the best numerical scores, our method produces comparable results. Lower scores for these objects are mainly due to our method's sensitivity to inaccurate silhouette masks provided by DiLiGenT-MV [21]. We project the GT surface onto the image plane and find up to 10-pixel inconsistency between the projected region and the GT mask. Thus, the silhouette loss Eq. (19) encourages our reconstructed surfaces to shrink to align with the smaller silhouettes.

Our method requires less effort for shape recovery than B-MVPS [21] and PS-NeRF [41]. While B-MVPS [21] calibrates 96 light directions and intensities, we use a self-calibrated PS method for input azimuth maps. PS-NeRF [41] uses 15 view $\times$ 96 light $=$ 1440 images to optimize multiple MLPs that model shape and appearance, which requires a high computational cost. It takes PS-NeRF [41] over 20 hours per object on an RTX 3090 GPU. In contrast, our approach optimizes a single MLP with 15 azimuth maps, taking approximately 3 hours per object on an RTX 2080Ti GPU.

Table 3 reports MAE for 5 test and all 20 viewpoints, and Fig. 8 visually compares recovered normal maps. MVAS improves normal accuracy compared to SDPS [5] and outperforms PS-NeRF [41] in 4 objects, demonstrating TSC's effectiveness in constraining surface normals from multi-view observations. Since TSC imposes a direct constraint on surface normals, it is more effective than modeling a rendering process as in PS-NeRF [41].

[2] IDR [42] refers to it as mask loss, but we prefer to use "silhouette loss" after shape-from-silhouette [20].

[3] Different strategies for computing CD yield different results to the original papers. UA-MVPS crops the invisible bottom face and uses mesh vertices [18]; PS-NeRF [41] samples 10000 points from the mesh surface.

Table 2. **(Top)** Chamfer distance (↓) and **(Bottom)** F-score (↑) [18, 19] of recovered geometry on DiLiGenT-MV benchmark [21].

	Bear	Buddha	Cow	Pot2	Reading	Average
R-MVPS [28]	1.070	0.397	0.440	1.504	0.561	0.794
B-MVPS [21]	**0.212**	**0.254**	**0.091**	0.201	**0.259**	**0.203**
UA-MVPS [18]	0.414	0.452	0.326	0.414	0.382	0.398
PS-NeRF [41]	0.260	0.314	0.287	0.254	0.352	0.293
MVAS (ours)	0.243	0.357	0.216	**0.197**	0.522	0.307
R-MVPS [28]	0.262	0.698	0.760	0.198	0.519	0.487
B-MVPS [21]	**0.958**	**0.902**	**0.986**	0.946	**0.914**	**0.941**
UA-MVPS [18]	0.707	0.669	0.798	0.731	0.762	0.733
PS-NeRF [41]	0.898	0.806	0.856	0.919	0.785	0.853
MVAS (ours)	0.909	0.754	0.907	**0.962**	0.546	0.816

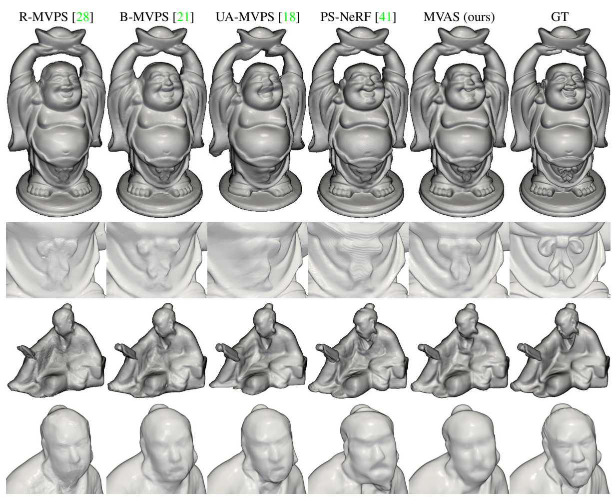


Figure 7. Visual comparison of recovered geometry. R-MVPS [28], B-MVPS [21], and UA-MVPS [18] require coarse geometry and use all 20 views for optimization, while PS-NeRF [41] and ours use a sphere initialization and 15 views.

Table 3. Mean angular error (↓) of recovered normal maps [21], evaluated using **(Top)** 5 test views and **(Bottom)** all 20 views.

Methods	# views	Bear	Buddha	Cow	Pot2	Reading	Average
R-MVPS [28]	5	12.80	13.67	10.81	14.99	11.71	12.80
B-MVPS [21]		3.80	10.57	**2.83**	5.76	**6.90**	**5.97**
PS-NeRF [41]		3.45	10.25	4.35	5.94	9.36	6.67
SDPS [5]		7.59	11.16	9.46	7.95	16.16	10.46
MVAS (ours)		**3.08**	**9.90**	3.72	**5.07**	10.02	6.36
R-MVPS [28]	20	12.70	13.63	10.92	14.91	11.79	12.79
B-MVPS [21]		3.81	10.58	**2.86**	5.72	**6.98**	**5.99**
PS-NeRF [41]		3.32	10.55	4.21	5.88	8.97	6.59
SDPS [5]		7.72	11.03	9.65	8.14	15.59	10.42
MVAS (ours)		**3.09**	**9.78**	3.74	**5.04**	10.06	6.34

4.2. MVAS for symmetric-light photometric stereo

Some photometric stereo methods can estimate azimuth angles well but struggle with zenith angles [3, 23]. This section shows how MVAS can be used for an uncalibrated photometric stereo setup to eliminate the need for tedious zenith estimation while allowing full surface reconstruction.

We use the setup shown in Fig. 9 to obtain multi-view azimuth maps. We place four lights symmetrically around the camera and the target object on a rotation table. In each view, we capture one ambient-light image and four lit images. The ambient-light images are used for SfM [30] to obtain the camera poses and are input to MVS [31] for comparison. Using the four lit images, The azimuth angles can be trivially computed from the ratio of the vertical to the horizontal difference image [23].

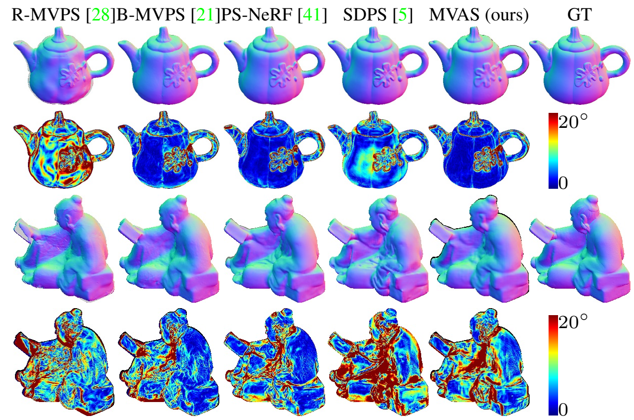


Figure 8. Visual comparison of recovered normal maps and angular error maps from the first view of DiLiGenT-MV [21] on the object “Pot2” and “Reading.”

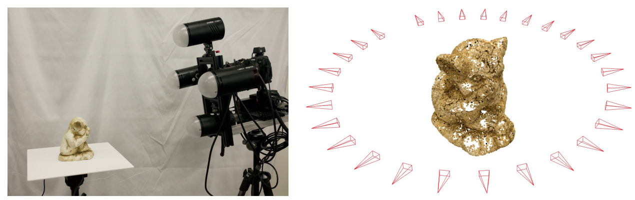


Figure 9. Our uncalibrated symmetric-light photometric stereo setup. Four lights are mounted symmetrically around the camera. We put the target object on a rotation table and capture about 30 views × 5 images in each view.

Figure 10 compares reconstructed surfaces and normals by Colmap [31] and MVAS. The first object shows a scene with challenging white planar faces. Photo-consistency-based MVS fails to recover the textureless region, while TSC succeeds in the planar region. This is possibly due to that TSC can still determine surface normals with wrong correspondences in a planar region, as discussed in Sec. 3.2. The second object has a dark surface, which is also challenging for photo-consistency, and Colmap [31] struggles to recover the correct surface normals.

4.3. MVAS with polarization imaging

This section shows the application of MVAS on azimuth maps obtained passively by a snapshot polarization camera, which makes the capture process as simple as MVS. Since

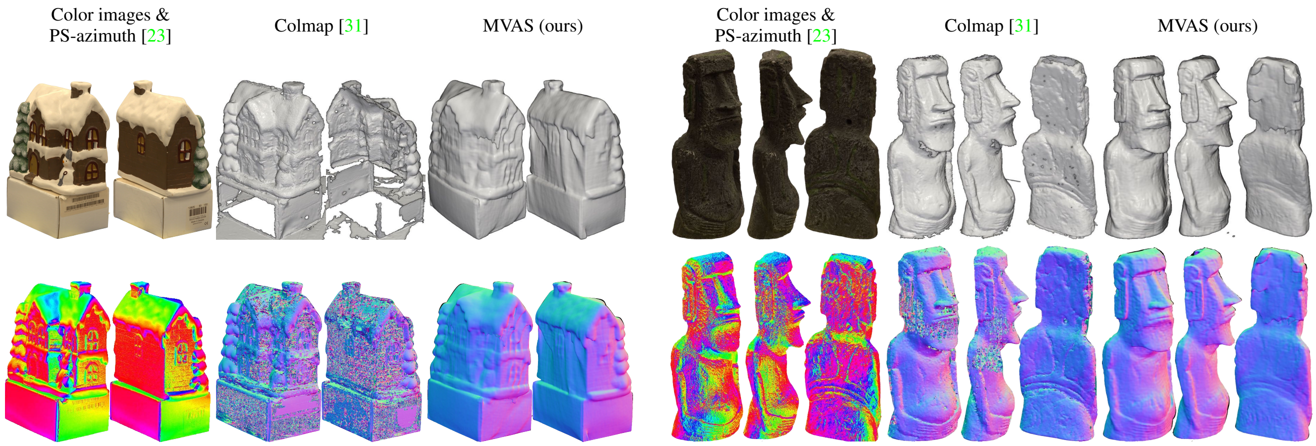


Figure 10. Visual comparison between MVAS and MVS on surface and normal reconstruction.

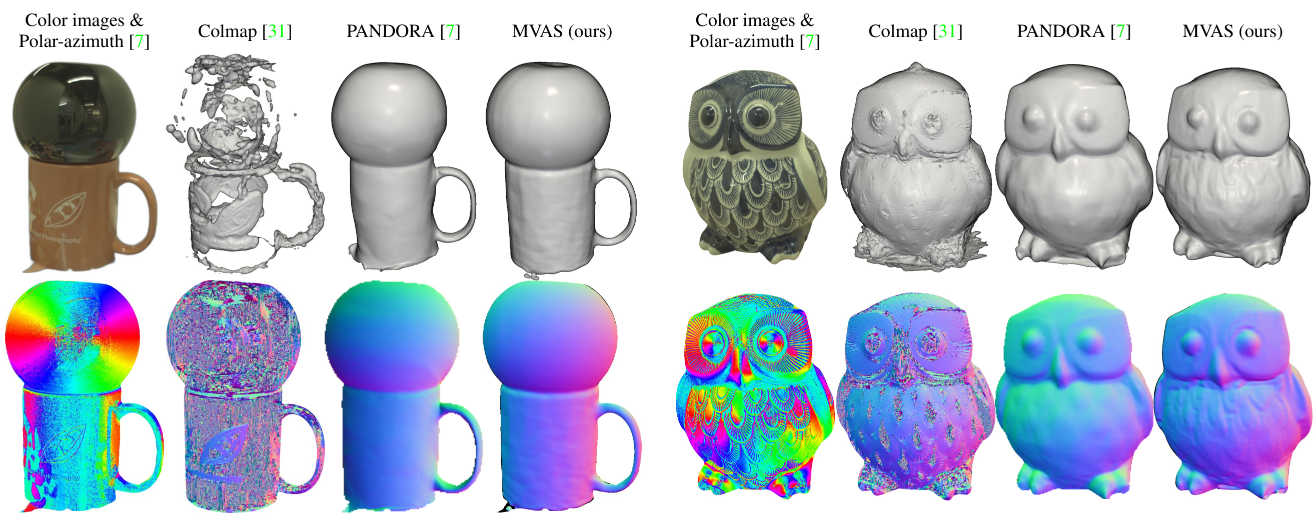


Figure 11. Qualitative comparison of recovered surfaces and normals using a polarization camera [7]. 35 views are used.

TSC is π-invariant, MVAS eliminates the need to correct the π-ambiguity [25]. Figure 11 compares the surface and normal reconstruction on the multi-view polarization image dataset [7]. We input the color images into Colmap [31] and reproduce the results of the polarimetric inverse rendering method PANDORA using their codes [7]. We modify our TSC loss to account for $\pm\frac{\pi}{2}$ ambiguity in polar-azimuth maps; see the supplementary material for details.

As shown in Fig. 11, MVS [31] breaks down for highly specular objects. Polar-azimuth observations are robust to such specularity and allow MVAS for faithful reconstruction. The comparison to PANDORA [7] shows that surfaces can be recovered without considering the degree of polarization or reflectance-light modeling.

5. Discussions

We present MVAS, an approach for reconstructing surfaces from multi-view azimuth maps. By establishing multi-view consistency in the tangent space and optimizing a neural SDF with the TSC loss, MVAS achieves comparable results to MVPS methods without zenith information. We verify MVAS's effectiveness with real-world azimuth maps obtained by symmetric-light photometric stereo and polarization measurements. Our results suggest that MVAS can enable high-fidelity reconstruction of shapes that have been challenging for traditional MVS methods.

Today, azimuth maps are still more expensive to obtain than ordinary color images, which may limit the application of MVAS. However, the situation will be changed when commercial polarimetric cameras are more accessible.

Acknowledgement

We thank Wenqi Yang, Akshat Dave, and Berk Kaya for code/data, and Boxin Shi, Min Li, and Heng Guo for discussions. This work was supported by JSPS KAKENHI Grant Number JP19H01123.

References

[1] Sony polarization image sensor. https://www.sony-semicon.com/en/products/is/industry/polarization.html. Accessed: 2023-03-23. 2

[2] Neil G Alldrin and David J Kriegman. Toward reconstructing surfaces with arbitrary isotropic reflectance: A stratified photometric stereo approach. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 1

[3] Manmohan Chandraker, Jiamin Bai, and Ravi Ramamoorthi. On differential photometric reconstruction for unknown, isotropic BRDFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(12):2941–2955, 2012. 1, 2, 7

[4] Ju Yong Chang, Kyoung Mu Lee, and Sang Uk Lee. Multiview normal field integration using level set methods. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2

[5] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K. Wong. SDPS-Net: Self-calibrating deep photometric stereo networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 7

[6] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[7] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. PANDORA: Polarization-aided neural decomposition of radiance. *Proc. of European Conference on Computer Vision (ECCV)*, 2022. 2, 8

[8] Yuqi Ding, Yu Ji, Mingyuan Zhou, Sing Bing Kang, and Jinwei Ye. Polarimetric Helmholtz stereopsis. In *Proc. of International Conference on Computer Vision (ICCV)*, 2021. 2

[9] Ondrej Drbohlav and Radim Sara. Unambiguous determination of shape from photometric stereo with unknown light sources. In *Proc. of International Conference on Computer Vision (ICCV)*, volume 1, pages 581–586, 2001. 2

[10] Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. Polarimetric normal stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 682–690, 2021. 2

[11] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

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

[13] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *Proc. of International Conference on Computer Vision (ICCV)*, 2007. 1

[14] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 6

[15] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(3):548–554, 2008. 1, 2

[16] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3D: High-quality depth sensing with polarization cues. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 3370–3378, 2015. 2

[17] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision (IJCV)*, 125(1):34–51, 2017. 2

[18] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 12601–12611, 2022. 2, 6, 7

[19] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 6, 7

[20] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1994. 6

[21] 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, 6, 7

[22] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1052–1061, 2019. 2

[23] Kazuma Minami, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Symmetric-light photometric stereo. In *Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2706–2714, 2022. 1, 2, 7, 8

[24] Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. Polarization-based transparent surface modeling from two views. In *Proc. of International Conference on Computer Vision (ICCV)*, volume 3, pages 1381–1381, 2003. 1

[25] Daisuke Miyazaki, Robby T Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *Proc. of International Conference on Computer Vision (ICCV)*, volume 3, pages 982–982, 2003. 8

[26] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics (TOG)*, 24(3):536–543, 2005. 1

[27] Stanley Osher, Ronald Fedkiw, and K Piechor. Level set methods and dynamic implicit surfaces. *Applied Mechanics Reviews*, 57(3):B15–B15, 2004. 5, 6

[28] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(8):1591–1604, 2016. 2, 6, 7

[29] Stefan Rahmann and Nikos Canterakis. Reconstruction of specular surfaces using polarization imaging. In *Proc. of*

Computer Vision and Pattern Recognition (CVPR), 2001. 1, 2

[30] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[31] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. of European Conference on Computer Vision (ECCV)*, 2016. 2, 7, 8

[32] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019. 1

[33] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:7462–7473, 2020. 6

[34] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Linear depth estimation from an uncalibrated, monocular polarisation image. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 109–125. Springer, 2016. 1, 2

[35] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(12):2875–2888, 2018. 2

[36] Christophe Stolz, Mathias Ferraton, and Fabrice Meriaudeau. Shape from polarization: A method for solving zenithal angle ambiguity. *Optics Letters*, 37(20):4218–4220, 2012. 1

[37] George Vogiatzis, Carlos Hernández Esteban, Philip HS Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(12):2241–2246, 2007. 1

[38] George Vogiatzis, Philip HS Torr, and Roberto Cipolla. Multi-view stereo via volumetric graph-cuts. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2005. 1

[39] Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(5):889–901, 2011. 1

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

[41] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. PS-NeRF: Neural inverse rendering for multi-view photometric stereo. In *Proc. of European Conference on Computer Vision (ECCV)*, 2022. 2, 6, 7

[42] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 5, 6

[43] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 85–102. Springer, 2020. 2

[44] Z. Zhou and P. Tan. Ring-light photometric stereo. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 265–279, 2010. 1

[45] Dizhong Zhu and William AP Smith. Depth from a polarisation+ RGB stereo pair. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7586–7595, 2019. 2

[46] Todd E Zickler, Peter N Belhumeur, and David J Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. *International Journal of Computer Vision (IJCV)*, 49(2):215–227, 2002. 5