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 $\pi$-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 \( \pi \)-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]. 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 \( \pi \) ambiguity. SfP studies determine surface normals by resolving this \( \pi \)-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 \( \pi \)-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 in-
3. Proposed Method

We aim to recover the shape from calibrated and masked azimuth maps. Let \( \Omega_i \) represent the \( i \)-th image pixel domain. For each view \( i \in \{1, 2, \ldots, C\} \), 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 \( P_i = [R_i, t_i] \in SE(3) \) and intrinsic perspective camera projection \( K_i \).

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

\[
R n = n^c. \tag{1}
\]

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

\[
n^c = \begin{bmatrix} n^c_x \\ n^c_y \\ n^c_z \end{bmatrix} = \begin{bmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{bmatrix}. \tag{2}
\]

From Eq. (2), we can derive the relation between \( n^c_x \) and \( n^c_y \) in terms of only the azimuth angle as

\[
n^c_x \sin \phi = n^c_y \cos \phi. \tag{3}
\]

We call \( t(\phi) \) 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),

\[
t^\top t = r_1 r_1 \sin^2 \phi + r_2 r_2 \cos^2 \phi - 2r_1 r_2 \cos \phi \sin \phi = \sin^2 \phi + \cos^2 \phi = 1,
\]

since \( r_1 \) and \( r_2 \) are orthonormal vectors.

The inset illustrates the second property. Let \( e_x, e_y \), and \( e_z \) be the unit direction vector of the \( x \)-, \( y \)-, and \( z \)-axis of the camera coordinates in the world coordinates. Then \( r_1 = e_x \) and \( r_2 = e_y \), which follows that \( t(\phi) \) 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,

\[
t \parallel n \times e_z. \tag{8}
\]

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 \( \pi \)-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 \( t_i(\mathbf{x}) = t(\phi_i(\Pi_i(\mathbf{x}))) \). By Eq. (6), a surface point \( \mathbf{x} \), its normal direction \( \mathbf{n} \), and its multi-view projected tangent vectors \( t \), should satisfy:

\[
\mathbf{n}(\mathbf{x})^\top t_i(\mathbf{x}) = 0 \quad \forall i.
\] (9)

Let \( \mathbf{T}(\mathbf{x}) = [t_1(\mathbf{x}), t_2(\mathbf{x}), ..., t_C(\mathbf{x})]^\top \in \mathbb{R}^{C \times 3} \) be the matrix formed by stacking projected tangent vectors of all \( C \) views. Then Eq. (9) reads

\[
\mathbf{T}(\mathbf{x})\mathbf{n}(\mathbf{x}) = 0.
\] (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 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 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., 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 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, 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.

| 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 $T(x)$ 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, $n(x)$ is the same and rank($T(x)$) is identically 2 for arbitrary correspondence. However, the normal direction of this plane can still be correctly determined in the case rank($T(x)$) = 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.

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

$$t(\phi + \pi) = -r_1 \sin \phi + r_2 \cos \phi = -t(\phi).$$

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

### 3.3. Multi-view azimuth stereo

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

$$\mathcal{J} = \iint_{\mathcal{M}} \sum_{i=1}^{C} \Phi_i(x) (n(x)^\top t_i(x))^2 \, d\mathcal{M}. \tag{12}$$

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

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

We can simplify Eq. (12) as follows:

$$\mathcal{J} = \iint_{\mathcal{M}} n^\top T n \, d\mathcal{M} \quad \text{with} \quad T = \sum_{i=1}^{C} \frac{\Phi_i t_i t_i^\top}{\sum_{i=1}^{C} \Phi_i}, \tag{14}$$

where we omit the dependence on the surface point $x$ 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 $f(x; \theta) : \mathbb{R}^d \rightarrow \mathbb{R}$, where $x \in \mathbb{R}^d$ is the 3D point coordinate, and $\theta \in \mathbb{R}^d$ are MLP parameters. The surface $\mathcal{M}$ is implicitly represented as the zero-level set of the SDF

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

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:

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

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 $X$. We evaluate the silhouette loss for pixels that do not have ray-surface intersections or are located outside the silhouette, denoted as $\bar{X}$.

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

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

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

$$n(x; \theta) = \nabla_x f(x; \theta). \tag{18}$$
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 $T(x)$ for the point $x$ from all visible views. First, we project the surface points onto all views and check their visibility in each, as shown in Fig. 5. 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

$$L_{\text{silhouette}} = \frac{1}{BM} \sum_{x \in X} \Psi(O(\Pi(x)), \sigma(\alpha f^*)),$$  \hspace{1cm} (19)

where $\Psi$ is the cross entropy function, and $\sigma(\cdot)$ is a sigmoid function with $\alpha$ 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]:

$$L_{\text{Eikonal}} = \mathbb{E}_X \left( \| \| n \|_2 - 1 \|^2 \right).$$ \hspace{1cm} (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 × 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].

<table>
<thead>
<tr>
<th></th>
<th>Bear</th>
<th>Buddha</th>
<th>Cow</th>
<th>Pot2</th>
<th>Reading</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-MVPS [28]</td>
<td>1.070</td>
<td>0.397</td>
<td>0.440</td>
<td>1.504</td>
<td>0.561</td>
<td>0.794</td>
</tr>
<tr>
<td>B-MVPS [21]</td>
<td>0.212</td>
<td>0.254</td>
<td>0.091</td>
<td>0.201</td>
<td>0.259</td>
<td>0.263</td>
</tr>
<tr>
<td>UA-MVPS [13]</td>
<td>0.414</td>
<td>0.452</td>
<td>0.326</td>
<td>0.414</td>
<td>0.382</td>
<td>0.398</td>
</tr>
<tr>
<td>PS-NeRF [41]</td>
<td>0.260</td>
<td>0.314</td>
<td>0.287</td>
<td>0.254</td>
<td>0.352</td>
<td>0.293</td>
</tr>
<tr>
<td>MVAS (ours)</td>
<td>0.243</td>
<td>0.357</td>
<td>0.216</td>
<td>0.197</td>
<td>0.522</td>
<td>0.307</td>
</tr>
<tr>
<td>R-MVPS [28]</td>
<td>0.262</td>
<td>0.698</td>
<td>0.760</td>
<td>0.198</td>
<td>0.519</td>
<td>0.487</td>
</tr>
<tr>
<td>B-MVPS [21]</td>
<td>0.958</td>
<td>0.902</td>
<td>0.986</td>
<td>0.946</td>
<td>0.914</td>
<td>0.941</td>
</tr>
<tr>
<td>UA-MVPS [18]</td>
<td>0.707</td>
<td>0.669</td>
<td>0.798</td>
<td>0.731</td>
<td>0.762</td>
<td>0.733</td>
</tr>
<tr>
<td>PS-NeRF [41]</td>
<td>0.898</td>
<td>0.806</td>
<td>0.856</td>
<td>0.919</td>
<td>0.785</td>
<td>0.853</td>
</tr>
<tr>
<td>MVAS (ours)</td>
<td>0.909</td>
<td>0.754</td>
<td>0.907</td>
<td>0.962</td>
<td>0.546</td>
<td>0.816</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th># views</th>
<th>Bear</th>
<th>Buddha</th>
<th>Cow</th>
<th>Pot2</th>
<th>Reading</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-MVPS [21]</td>
<td>5.80</td>
<td>10.57</td>
<td>2.83</td>
<td>5.76</td>
<td>6.90</td>
<td>5.97</td>
<td></td>
</tr>
<tr>
<td>PS-NeRF [41]</td>
<td>3.45</td>
<td>10.25</td>
<td>4.35</td>
<td>5.94</td>
<td>9.36</td>
<td>6.67</td>
<td></td>
</tr>
<tr>
<td>MVAS (ours)</td>
<td>3.08</td>
<td>9.90</td>
<td>3.72</td>
<td>5.07</td>
<td>10.02</td>
<td>6.36</td>
<td></td>
</tr>
<tr>
<td>B-MVPS [21]</td>
<td>3.81</td>
<td>10.58</td>
<td>2.86</td>
<td>5.72</td>
<td>6.98</td>
<td>5.99</td>
<td></td>
</tr>
<tr>
<td>PS-NeRF [41]</td>
<td>3.32</td>
<td>10.55</td>
<td>4.21</td>
<td>5.88</td>
<td>8.97</td>
<td>6.59</td>
<td></td>
</tr>
<tr>
<td>MVAS (ours)</td>
<td>3.09</td>
<td>9.78</td>
<td>3.74</td>
<td>5.04</td>
<td>10.06</td>
<td>6.34</td>
<td></td>
</tr>
</tbody>
</table>

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 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
TSC is $\pi$-invariant, MVAS eliminates the need to correct the $\pi$-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


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


