Sparse Views, Near Light:

A Practical Paradigm for Uncalibrated Point-light Photometric Stereo

Mohammed Brahimi[1,2]   Bjoern Haefner[1,2,4][*]   Zhenzhang Ye[1,2]   Bastian Goldluecke[3]   Daniel Cremers[1,2]

[1] TU Munich, [2] Munich Center for Machine Learning, [3] University of Konstanz, [4] NVIDIA

{mohammed.brahimi, bjoern.haefner, yez, cremers}@tum.de

bastian.goldluecke@uni-konstanz.de

Abstract

Neural approaches have shown a significant progress on camera-based reconstruction. But they require either a fairly dense sampling of the viewing sphere, or pre-training on an existing dataset, thereby limiting their generalizability. In contrast, photometric stereo (PS) approaches have shown great potential for achieving high-quality reconstruction under sparse viewpoints. Yet, they are impractical because they typically require tedious laboratory conditions, are restricted to dark rooms, and often multi-staged, making them subject to accumulated errors. To address these shortcomings, we propose an end-to-end uncalibrated multi-view PS framework for reconstructing high-resolution shapes acquired from sparse viewpoints in a real-world environment. We relax the dark room assumption, and allow a combination of static ambient lighting and dynamic near LED lighting, thereby enabling easy data capture outside the lab. Experimental validation confirms that it outperforms existing baseline approaches in the regime of sparse viewpoints by a large margin. This allows to bring high-accuracy 3D reconstruction from the dark room to the real world, while maintaining a reasonable data capture complexity.

1. Introduction

The challenge of 3D reconstruction stands as a cornerstone in both computer vision and computer graphics. Despite notable progress in recovering an object's shape from dense image viewpoints, predicting consistent geometry from sparse viewpoints remains a difficult task. Contemporary approaches employing neural data structures depend heavily on extensive training data to achieve generalization in the context of sparse views. Additionally, the presence of a wide baseline or textureless objects form significant obstacles. In contrast, photometric methodologies like photometric stereo (PS) excel in reconstructing geometry, even in textureless regions. This capability is attributed to the abun-
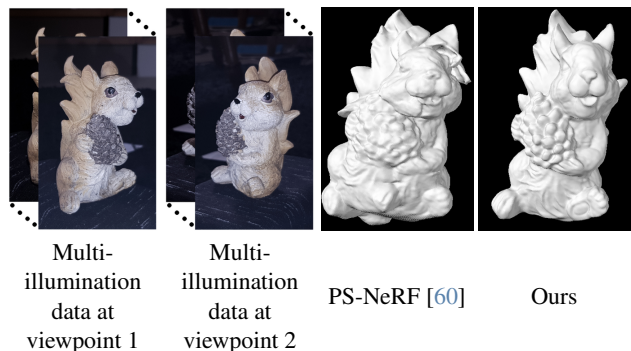

Figure 1. We introduce the first framework for multi-view uncalibrated point-light photometric stereo. Given a set of PS images captured from different viewpoints (left), our method recovers high-fidelity 3D reconstruction (right). The acquisition of uncalibrated point-light PS imagery captured under ambient lighting in a sparse multi-view setup does not only allow for easy data capture, but also leads to 3D reconstructions of unprecedented detail. Here, with as few as two views we are able to reconstruct the squirrel's 3D geometry at higher precision than the state-of-the-art.

dance of shading information derived from images acquired under diverse illumination. Yet, such approaches typically require a controlled laboratory setup to fulfill the necessary dark room and directional light assumptions. As a consequence, PS approaches become impractical beyond the confines of a laboratory.

To address these shortcomings we combine state-of-the-art volume rendering formulations with a sparse multi-view photometric stereo model. In particular, we advocate a physically realistic lighting model that combines ambient light and uncalibrated point-light illumination. Our approach facilitates simplified data acquisition, and we introduce a novel pipeline capable of reconstructing an object's geometry from a sparse set of viewpoints, even if the object is completely textureless. Furthermore, since we assume a point-light model instead of distant directional lighting, we can infer absolute depth from a single viewpoint, allowing us to address the challenge of wide camera baselines.

In detail, our contributions are as follows:

- We introduce the first framework for multi-view uncalibrated point-light photometric stereo. It combines a state-of-the-art volume rendering formulation with a physically realistic model of ambient light and point lights.
- We eliminate the need for a dark room environment, dense capturing process, and distant lighting. Thereby we enhance the accessibility and simplify data acquisition for setups beyond traditional laboratory settings.
- We validate that it can be successfully used for accurate shape reconstruction of textureless objects in highly sparse scenarios with wide baselines.
- Despite the absence of pre-processing or vast training data, we outperform cutting-edge approaches that either rely only on static ambient illumination or PS imagery.

2. Related Work

In this section, we categorize pertinent works into two distinct groups. The first category focuses on neural 3D reconstruction, emphasizing the use of images captured under static ambient illumination. Subsequently, the second category delves into photometric stereo (PS) approaches, which specifically involve methods assuming images captured under varying illumination conditions.

2.1. Neural 3D reconstruction

In our context, the domain of neural 3D reconstruction under static ambient light can be bifurcated into two distinct subcategories. The first comprises methodologies that presume a dense sampling of the viewing sphere [5, 11, 12, 14, 36, 38, 39, 52, 63, 64, 68]. Conversely, the second subcategory encompasses techniques adept at handling sparse camera viewpoints [7, 9, 13, 16, 22, 33, 43, 47, 51, 53, 56–59, 61, 62, 65–67].

Given our emphasis on a sparse set of camera viewpoints, our focus will be directed toward the latter set of methodologies. The majority of sparse reconstruction approaches within this category hinge on training with a specific dataset [9, 13, 16, 47, 53, 57–59, 61, 62, 65, 67]. The reliance on extensive training data renders these techniques susceptible to errors when confronted with objects not represented in the training dataset, thereby introducing sensitivity to the domain gap between training and test data.

Efforts have been undertaken to mitigate the challenges associated with the domain gap. One approach involves fine-tuning pre-trained architectures during test time with sparse input data [22, 33, 43]. However, while fine-tuning enables fine-scale refinement of the predicted reconstruction, the broader issue of generalization to entirely different datasets persists as a challenge. Furthermore, it is crucial to note that all the previously mentioned sparse approaches presume small camera baselines and textured objects. To achieve a full reconstruction of an object from sparse views, it becomes imperative to develop texture-

agnostic approaches capable of handling wide baselines.

Another set of approaches involves the utilization of pre-training to establish a geometric prior [51, 56, 66]. Yu *et al*. [66] leverages this geometric prior to predict a depth and normal map, albeit with a limitation in fine-scale details due to its coarse resolution. Notably, [51, 56] address this limitation. Wu *et al*. [56] relax the assumption on small baselines, although a reasonable overlap of sparse views remains necessary for consistent 3D reconstruction. Vora *et al*. [51] demonstrate promising results with much less overlap, employing only three viewpoints. However, the use of template shapes as a prior term introduces the risk of the reconstruction quality being affected by the domain gap curse.

In contrast, our approach eliminates the requirement for training data and operates effectively in wide baseline scenarios, independent of the object's texture. This accomplishment can be attributed to the integration of PS principles, which we will elucidate in the subsequent discussion.

2.2. Photometric Stereo

The principle of PS [54] has a long and established history, renowned for yielding high-quality and finely detailed geometric estimates, due to its capturing process: multiple images under different illumination of a static object are captured from a single camera viewpoint.

In our case, PS approaches can thus be categorized into two main types. The first category pertains to single-view methods [17, 23, 29, 30, 32, 35, 40, 45, 46, 50], wherein data is acquired as described above. The second category involves multi-view methods [1, 10, 21, 26–28, 31, 41, 55, 60, 72], where the capturing process is repeated at multiple camera viewpoints.

Within the context of our task, we concentrate on the multi-view scenario. Many existing approaches in this domain assume knowledge of the light source for each image, earning them the designation of calibrated PS (CPS). In this case, frequently the light is parameterized as directional [1, 10, 28, 72], while approaches employing point-light illumination are more sparsely represented [10, 28, 31], owing to the intricate model and the much more complex calibration process [46]. In a broader context, the calibration of each image's light source is impractical and tedious, underscoring the advantage of the uncalibrated case, which inherently facilitates a more straightforward and streamlined capturing process.

Efforts have been dedicated to exploring the uncalibrated PS (UPS) scenario, where there is no prior knowledge of the light source [21, 26, 27, 41, 55, 60]. Nevertheless, all these uncalibrated methods typically assume a distant and directional light model or permit a lighting model based on spherical harmonics [2, 4, 18, 19, 44, 48, 49], albeit at the expense of being restricted to Lambertian objects [55]. Consequently, the exploration of the uncalibrated point-

light case remains uncharted territory, particularly in the multi-view scenario.

Despite the impractical assumption of calibrated light or the constraint of uncalibrated directional lighting, the aforementioned multi-view photometric stereo approaches demand a dark room and lack accommodation for ambient light [1, 10, 21, 26–28, 31, 41, 60, 72]. Additionally, they comprise multiple stages, rendering them prone to initialization and accumulation errors [21, 26–28, 31, 41, 55, 60]. Furthermore, these methods assume Lambertian objects [31, 55] and crucially rely on a dense sampling of camera viewpoints around the object [21, 28, 31, 41, 55].

The approach most closely related to ours is the sparse method PS-NeRF [60], which assumes uncalibrated directional lighting with imagery captured in a dark room. This method comprises multiple stages, where initially, a pretrained network [8] is utilized to estimate the initial light, along with normal information for each image and camera view. Subsequently, this acquired information is employed to estimate the object's mesh through the optimization of an occupancy field [39]. The outputs of both stages are then combined to derive a consistent normal map of the object. Although PS-NeRF demonstrates impressive results in a sparse scenario with five viewpoints, its susceptibility to error accumulation is a potential limitation, given the involvement of multiple stages.

Our proposed model addresses the aforementioned limitations comprehensively. Notably, we excel in handling sparse yet wide baseline scenarios, where existing state-of-the-art methods encounter challenges. Our investigation tackles the open challenge of multi-view uncalibrated point-light PS within an end-to-end framework, liberating us from the constraints associated with Lambertian objects. Furthermore, our approach facilitates easy data capture in the presence of ambient light, eliminating the need for a dark room.

3. Setting and image formation model

We consider an object under static illumination photographed from several different viewpoints. For every viewpoint, we are given a set of $N+1$ images $(I^l)_{l=0,\ldots,N}$. Image I^0 is captured only under static ambient illumination and called the ambient image. For images I^1 to I^N, the object is in addition to the ambient illumination also illuminated by a single achromatic point-light source with scalar light intensity $L_0 \in \mathbb{R}^+$, placed at a different location $\mathbf{p}^l \in \mathbb{R}^3$ for each image and viewpoint. We use the common assumption in volume rendering [36, 52, 56, 64] that the image brightness I^l is the same as the accumulated radiance of the volume rendering, see Eq. (4). Note that we do not add an index for the viewpoint to the image to not clutter notation, whenever we later evaluate an image at a pixel location, we assume it implicitly means a pixel at a specific viewpoint. In particular, a sum over all pixels is the sum over all pixels at every viewpoint location.

According to the linearity of the radiance, the total radiance for the image l is given by a sum $\mathcal{L}^l = \mathcal{L}_{\text{amb}} + \mathcal{L}_{\text{pt}}^l$, where $\mathcal{L}_{\text{amb}}$ is the radiance due to the static illumination, the same for each image, and $\mathcal{L}_{\text{pt}}^l$ is the radiance due to the point-light source used for that particular image. We will first express the radiance due to the point-light source with respect to the material and lighting parameters, which allows to take into account the different illuminations of the images. Subsequently, we elaborate on how we deal with the ambient component.

3.1. Point-light illumination and material

The field $L(\mathbf{x}, \mathbf{n}, \mathbf{v})$ maps points $\mathbf{x} \in \mathbb{R}^3$, normal direction $\mathbf{n}(\mathbf{x}) \in \mathbb{S}^2$ and viewing direction $\mathbf{v} \in \mathbb{S}^2$ to its radiance. It can be expressed in terms of geometry, material and lighting with the classical rendering equation [24]. Since for now we only consider a single achromatic point-light source with intensity L_0 at position $\mathbf{p}$, this equation simplifies to

$$L(\mathbf{x}, \mathbf{n}, \mathbf{v}) = \frac{L_0}{\|\mathbf{p} - \mathbf{x}\|^2} f_{\text{r}}(\mathbf{x}, \mathbf{n}, \mathbf{v}, \mathbf{l}) \max(0, \mathbf{n} \cdot \mathbf{l}), \quad (1)$$

where $\mathbf{l} = \frac{\mathbf{p} - \mathbf{x}}{\|\mathbf{p} - \mathbf{x}\|}$.

As a model for the spatially varying reflectance f_{r}, we use the simplified Disney BRDF [25]. It is relatively simple yet expressive enough to resemble a wide variety of materials. It has already been successfully used in several other prior works [5, 34, 69]. The Disney BRDF requires three parameters to be available at every point $\mathbf{x} \in \mathbb{R}^3$: An RGB diffuse albedo $\rho \in \mathbb{R}^3_{\geq 0}$, a roughness $\alpha_{\text{r}} > 0$ and a specular albedo $\alpha_{\text{s}} \in [0, 1]$.

As in [5], all three BRDF parameters can be represented with MLPs. One MLP with network parameters γ_1 computes the diffuse parameter ρ, a second MLP with network parameters γ_2 computes the specular parameters $(\alpha_{\text{r}}, \alpha_{\text{s}})$. To encompass all point-light and BRDF parameters, we concatenate them into the vectors $\phi = (\mathbf{p}^1, \ldots, \mathbf{p}^M)$, which includes all light source positions (where M is the product of N with the number of viewpoints), and $\gamma = (\gamma_1, \gamma_2)$, representing the BRDF parameters. Both ϕ and γ serve as the network parameters to be learned during optimization. However, note that one of our contributions is a novel strategy to deal with the diffuse albedo later on in Sec. 5.2, which dramatically improves performance in extremely sparse scenarios.

We assume that the intensity L_0 remains constant across *all* images and views. This assumption is reasonable, *e.g.*, when using a smartphone's flashlight or a sole LED, as shown in our experiments. A constant intensity introduces a scale ambiguity between the f_{r} and L_0. We set $L_0 \equiv 1$ and subsequently estimate a scaled version of the BRDF, *i.e.*, the intensity is absorbed into the BRDF. This is feasible, since we can scale the Disney BRDF arbitrarily.

3.2. Ambient light

Given the provided ambient image, we adopt the approach from [71], which involves subtracting the ambient image from all other images. This process effectively eliminates ambient radiance from our considerations, simulating a traditional dark room scenario. A drawback is that in practice, the captured images contain additive noise, hence such a subtraction will double the variance of the noise distribution, resulting in noisier input images. Nevertheless, as can be seen in the results, our approach works successfully with this strategy even if we use a simple smartphone camera for the data capture.

However, this simplification is enabled by the fact that the point-light sources are near the object, making it easy to obtain a well exposed contribution from the point-light source. In methods assuming directional lighting, the practical simulation often involves using a point-light source placed far from the object, ensuring that the object's diameter is significantly smaller than the distance between the point-light and the object. For example, for the DiLiGenT-MV dataset [28], the light sources are roughly 1m from the object, whereas for our setup, the distance is approximately 25cm. As a consequence, for a given static ambient illumination, the distant point-light will need to be $16\times$ brighter in the Diligent setup to obtain a similar signal to noise ratio. This increased brightness requirement escalates the hardware demands, especially if the highest quality is sought.

3.3. Other illumination effects

We do not model indirect lighting and cast-shadows explicitly. However, multiple previous works [3, 20, 34, 37, 70] and PS frameworks [10, 17, 28, 46] have demonstrated that satisfactory results can be achieved without taking these into account. In addition, we use the robust L^1-norm in our data term (7) to further increase robustness. Furthermore, we will expose a simple strategy in Sec. 5 to reduce the impact of cast-shadows, which inevitably occur when capturing a nonconvex object under changing illumination.

4. Volume rendering of the surface

For differentiable rendering we build upon VolSDF [64] because in contrast to NeRF [36] it allows us to decouple geometric representation and appearance. As discussed in [47, 56, 65], VolSDF exhibits poor reconstruction quality in the sparse viewpoint scenario due to the shape-radiance ambiguity. We eliminate this ambiguity thanks to photometric stereo, providing shading cues from different illuminations for each viewpoint. To this end, we employ the radiance field in Eq. (1) and apply it to the multi-view uncalibrated point-light PS setting. The integration of differentiable volume rendering with a physically realistic lighting model provides an efficient solution to the multi-view

photometric stereo problem, enabling high-quality reconstruction even with sparse viewpoints, as demonstrated in the subsequent results.

Geometry model and associated density. The geometry of the scene is modeled with a signed distance function (SDF) $d : \Omega \to \mathbb{R}$ within the volume $\Omega \subset \mathbb{R}^3$. We follow [42, 52, 63, 64] and represent the SDF with an MLP $d(\mathbf{x}; \theta)$ parametrized by θ. Note that we assume positive values of the SDF d inside the object, thus the normal vector is given by $\mathbf{n} = \nabla d / \|\nabla d\|$ on Ω. To render the scene, we follow [64] and transform the SDF into a density $\sigma : \Omega \to \mathbb{R}_{\geq 0}$ by means of the transformation

$$\sigma(\mathbf{x}) = \beta^{-1} \Psi_\beta(-d(\mathbf{x})), \qquad (2)$$

where Ψ_β denotes the cumulative distribution function of the Laplace distribution with zero mean. The positive scale parameter β is learned during optimization. For simplicity, we assume it is included in the set of SDF parameters θ.

Volume Rendering. Using the density (2), we can now set up a distribution of weights $w(t)$ along each ray $\mathbf{x}(t) = \mathbf{c} + t\mathbf{v}, t \geq 0$ defined by the center of projection $\mathbf{c}$ of the camera and viewing direction $\mathbf{v}$,

$$w(t) = \sigma(\mathbf{x}(t)) \exp\left(-\int_0^t \sigma(\mathbf{x}(s)) \, \mathrm{d}s\right). \qquad (3)$$

This is the derivative of the opacity function, which is monotonically increasing from zero to one, and thus w is indeed a probability distribution [64]. In particular, the integral

$$\mathcal{L}_p = \int_0^\infty w(t) L(\mathbf{x}(t), \mathbf{n}(t), \mathbf{v}) \, \mathrm{d}t, \qquad (4)$$

to compute radiance $\mathcal{L}_p$ in the pixel $p \in \mathbb{R}^2$ corresponding to the ray $\mathbf{x}(t)$ is well-defined. Intuitively, we accumulate the radiance field (1) of the point-light source at locations where opacity increases, i.e. where we move towards the inside of the object. For an ideal sharp boundary, w would be a Dirac distribution with its peak placed at the boundary.

As was done in [64] we approximate the integral (4) using the quadrature rule

$$\mathcal{L}_p \approx \sum_{i=1}^{m-1} (t_{i+1} - t_i) w(t_i) L(\mathbf{x}(t_i), \mathbf{n}(t_i), \mathbf{v}). \qquad (5)$$

for numerical integration with a discrete set of samples $t_1 < t_2 < \cdots < t_m$. Note that the integral (3) required for $w(t_i)$ in (5) can be numerically computed using an analogous formula while iteratively accumulating the sum (5).

5. Training objective and optimization

In order to reconstruct the scene by inverse rendering, we train our neural networks from the available input images. We optimize for the geometry parameters θ of the SDF network and point-light and BRDF parameters ϕ and γ of the appearance network, respectively.

5.1. Vanilla loss function with MLP for albedo

The overall loss consists of the sum of an inverse rendering loss $E_{\mathrm{RGB}}(\theta, \phi, \gamma)$, an Eikonal loss $E_{\mathrm{eik}}(\theta)$ and a mask loss $E_{\mathrm{mask}}(\theta)$ for the SDF,

$$E(\theta, \phi, \gamma) = E_{\mathrm{RGB}}(\theta, \phi, \gamma) + \lambda_1 E_{\mathrm{eik}}(\theta) + \lambda_2 E_{\mathrm{mask}}(\theta), \quad (6)$$

with respective weighting parameters $\lambda_1, \lambda_2 \geq 0$ given as hyperparameters. We will describe the different terms in detail in the following.

The **rendering loss**

$$E_{\mathrm{RGB}}(\theta, \phi, \gamma) = \sum_p \sum_{l \in B_\tau(p)} \left\| I_p^l - \mathcal{L}_p^l(\theta, \phi, \gamma) \right\|_1 \quad (7)$$

encourages that the rendered images resemble the input images. As a strategy to deal with cast-shadows, for each pixel p, we use only the τ brightest views $B_\tau(p) \subset \{1 \dots N\}$, where $\tau > 0$ is a hyper-parameter. This reduces the impact of cast-shadows, as a pixel will have the smallest possible brightness when it lies in shadow. Such a strategy is necessary since despite the use of an L^1-norm to improve robustness against outliers, some regions with frequent cast-shadows would still exhibit small artefacts, as can be seen in Fig. 2. In order to avoid wasting meaningful information due to this heuristic, we set $\tau = N$ for the first half of the iterations, and then set it to $\tau = \lceil \frac{3}{4} N \rceil$ for the second half for improved handling of cast-shadows.

The **Eikonal term**

$$E_{\mathrm{eik}}(\theta) = \sum_{\mathbf{x}} (\|\nabla_{\mathbf{x}} d(\mathbf{x}; \theta)\| - 1)^2 \quad (8)$$

encourages $d(\mathbf{x}; \theta)$ to approximate an SDF, as it penalizes deviations from the respective partial differential equation which every SDF satisfies, see e.g. [15].

Finally, we follow [5, 52] and employ the **mask loss**

$$E_{\mathrm{mask}}(\theta) = \sum_p \mathrm{BCE}(M_p, W_p(\theta)), \quad (9)$$

to impose silhouette consistency by means of the binary cross entropy (BCE) between the given binary mask M_p at the pixel p and the sum $W_p(\theta) = \sum_{i=1}^{m-1} w(t_i; \theta)$ of weights at the sampling locations t_i used in (5).

5.2. Using the optimal diffuse albedo

In this section, we introduce a strategy to handle the diffuse albedo. This is inspired by [17], however in their single-view approach they are only able to handle the least-squares case, *i.e.* L^2, while we extend their formulation to objective functions of the form of an L^p-norm, with $p = 1$ in our case, see Eq. (7). In lieu of employing an MLP for modeling the diffuse albedo, we opt for approximating an optimal solution considering the remaining parameters. This allows us to exclude the diffuse albedo from the optimization
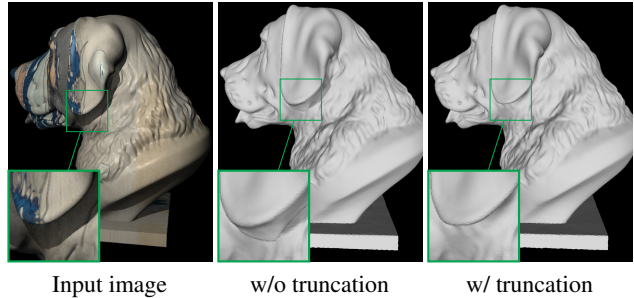

Input image   w/o truncation   w/ truncation

Figure 2. Effect of cast-shadows on the result. The region behind the ear exhibits frequent cast-shadows in the input images (left), which induce artefacts in the estimated mesh (middle). Our truncation strategy (Sec. 5.1) successfully eliminates these (right).

process. Our subsequent demonstration will underscore the significant benefits of this approach in the context of sparse views, leading to markedly improved final geometry. A key technical change is the domain of the diffuse parameter ρ. So far, it has been a vector field defined in world space Ω, implemented with a neural network. However, in the following we only try to recover the projection of ρ on the image planes. Thus, the input to ρ will be a pixel, and the output is the diffuse albedo at the intersection between the pixel's ray and the object surface. In particular, the set of BRDF parameters is reduced to $\gamma = \gamma_2$ and now only includes specular parameters.

Reformulation of the data term. We will now derive an expression for the data term which allows to solve for diffuse albedo ρ when given all of the other parameters. Note that the SVBRDF can be decomposed into a diffuse part and a specular part [25],

$$f_{\mathrm{r}}(\mathbf{x}, \mathbf{n}, \mathbf{v}, \mathbf{l}) = \frac{\rho(\mathbf{x})}{\pi} + f_{\mathrm{s}}(\mathbf{x}, \mathbf{n}, \mathbf{v}, \mathbf{l}), \quad (10)$$

where the specular part f_{s} depends on both the roughness and specular albedo. If we denote the BRDF-independent factor of the radiance field (1) by

$$S^l(\mathbf{x}, \mathbf{n}) = \frac{L_0}{\|\mathbf{x} - \mathbf{p}^l\|^2} \max(0, \mathbf{n} \cdot \mathbf{l}), \quad (11)$$

this implies that the total radiance field is the sum of diffuse radiance $\frac{\rho}{\pi} S^l$ and specular radiance $L_s^l = f_{\mathrm{s}} S^l$.

Inserting this into the volume rendering equation (4), we can thus re-arrange the data term (7) as

$$E_{\mathrm{RGB}}(\theta, \phi, \gamma, \rho) = \sum_p \sum_{l \in S_\tau(p)} \left\| b_p^l(\theta, \phi, \gamma) - \rho_p a_p^l(\theta, \phi) \right\|_1$$

$$\text{with } a_p^l(\theta, \phi) = \frac{1}{\pi} \int_0^\infty w_\theta(t) S^l(\mathbf{x}(t), \mathbf{n}_\theta(t)) \, \mathrm{d}t,$$

$$b_p^l(\theta, \phi, \gamma) = I_p^l - \int_0^\infty w_\theta(t) L_s^l(\mathbf{x}(t), \mathbf{n}_\theta(t), \mathbf{v}) \, \mathrm{d}t.$$

$$(12)$$

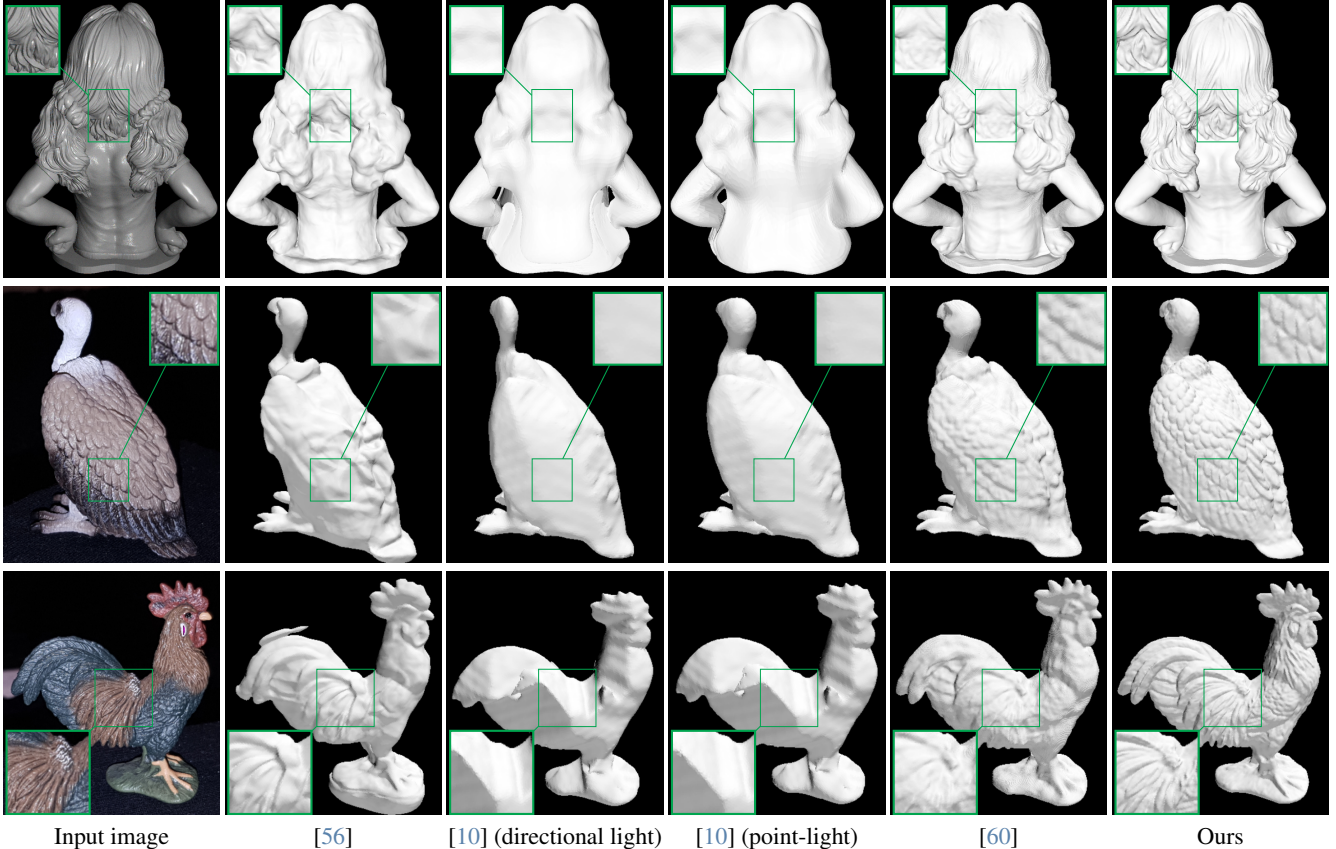| Input image | [56] | [10] (directional light) | [10] (point-light) | [60] | Ours |

Figure 3. Full 3D reconstruction of one synthetic and two real objects from 6 viewpoints. While some of the baseline reconstructions (with directional or point-light models) exhibit texture-induced patterns, distortions or other artefacts, the proposed framework consistently yields artefact- and distortion-free reconstructions with finer geometric details. Additional results are shown in the supplementary.

When keeping θ, ϕ, and γ fixed, this expression can be minimized pixel-wise for ρ as a linear least absolute deviation problem, as we show next.

Alternate treatment of the diffuse albedo. It is now possible to completely exclude the diffuse albedo ρ from the optimization and simply replace it by the optimal value

$$\rho^*_{\theta,\gamma} = \arg\min_{\rho} E_{\text{RGB}}(\theta, \phi, \gamma, \rho) \qquad (13)$$

with respect to the other parameters. In our case, we are facing a linear least absolute deviation problem which can be solved individually for each pixel.

We thus approximate the solution using a few iterations of iterative reweighted least squares [6], as it efficiently solves the problem while providing a differentiable expression. In order to simplify notation, we omit the dependency w.r.t. the parameters and denote only the channel-wise operations, where we iteratively compute

$$\rho^*_p \approx \rho^{(k)}_p = \frac{b_p^\top W_p^{(k)} a_p}{a_p^\top W_p^{(k)} a_p} \qquad (14)$$

with diagonal matrices

$$W_p^{(k)} = \begin{cases} \text{diag}(\max(\epsilon, |b_p - \rho_p^{(k-1)} a_p|)^{-1}) & \text{if } k \geq 1, \\ I & \text{otherwise.} \end{cases} \qquad (15)$$

Above, k is the index of iteration, and a_p and b_p are the vectors whose components are given in Eq. (12). The small constant $\epsilon > 0$ avoids numerical issues, and I is an identity matrix of matching size. The gradient of $\rho_p^{(k)}$ with respect to θ and γ is computed using automatic differentiation, however the gradient of $W_p^{(k)}$ is not back-propagated, as it makes the optimization much more difficult and degrades the quality of the results.

In the forthcoming evaluation section, we demonstrate that this formulation delivers exceptional performance in 3D reconstruction from sparse viewpoints. Notably, it proves particularly advantageous in the most sparse scenarios, where only two viewpoints are available, showcasing its superiority over the previous MLP-based formulation.

	dog1			

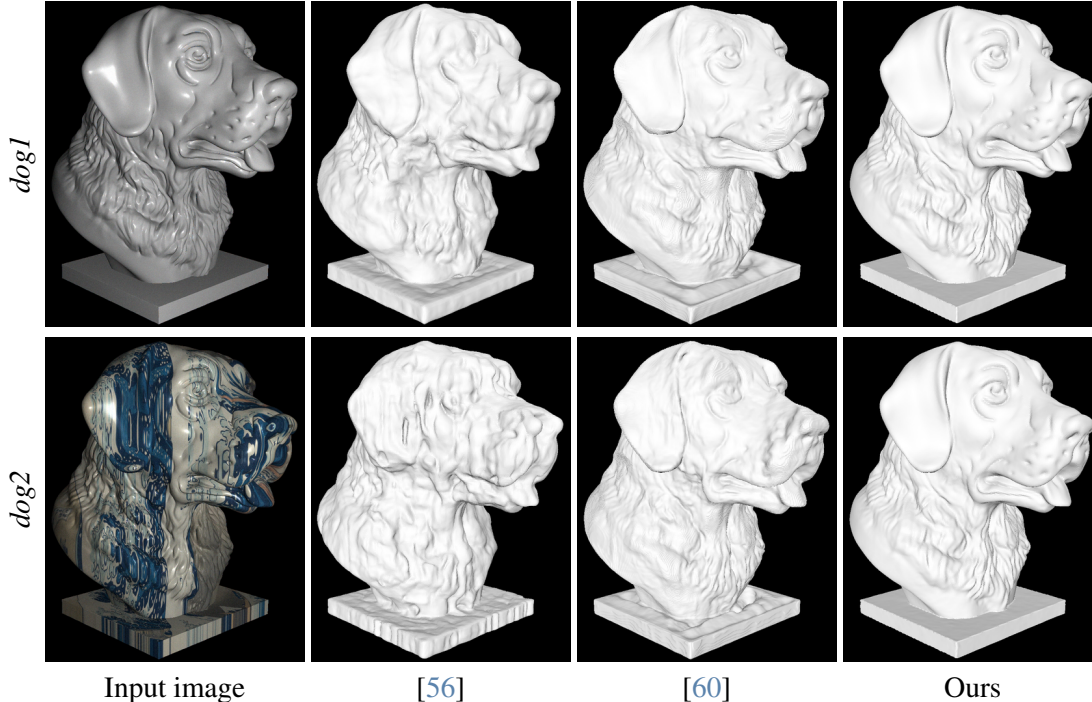Input image   [56]   [60]   Ours

Figure 4. Results for both *dog1* and *dog2* obtained with 5 viewpoints reveal a notable performance degradation in [56] and [60] when confronted with complex textures. This deterioration is attributed to a generalization failure of their learned priors. In contrast, our approach, which does not rely on any prior, provides a superior reconstruction that is unaffected by the complex texture.

	↓RMSE×100					↓MAE				
	[56]	[10]D	[10]P	[60]	Ours	[56]	[10]D	[10]P	[60]	Ours
dog1	2.1	8.9	13.9	2.1	**0.4**	22.0	39.9	39.2	18.8	**4.6**
dog2	2.2	8.4	13.4	2.1	**0.4**	23.2	29.6	40.3	19.5	**4.8**
girl1	0.9	6.1	2.3	1.5	**0.3**	17.2	41.8	46.5	23.8	**9.6**
girl2	1.5	5.2	4.8	1.5	**0.3**	23.9	40.0	47.3	23.4	**10.2**

Table 1. RMSE and MAE for full 3D reconstructions. RMSE is computed based on the vertex to mesh distance, and the MAE is computed using the angular error between the normals of a vertex and its closest point in the ground truth mesh. *OurAlbedoNet* leads to highly similar results as our main framework in this scenario.

6. Results

To substantiate the efficacy of our framework, we conduct evaluations on both synthetic and real-world datasets. We generate four synthetic scans by combining two distinct geometries with two different materials. The first material is white, rendering it textureless, while the second material exhibits a high degree of texture. This design enables us to quantify the influence of texture on the obtained results. The four synthetic scans are denoted as *dog1*, *dog2*, *girl1*, and *girl2*. Additionally, we acquire real-world data for six objects, namely, *bird*, *squirrel*, *hawk*, *rooster*, *flamingo* and *pumpkin*. We also consider the DiLiGenT-MV dataset [28].

Evaluation. We first evaluate full 3D reconstruction using only six viewpoints in total, except for *dog1* and *dog2* where five viewpoints are considered. We evaluate against both state-of-the-art sparse viewpoint reconstruction approaches

assuming static illumination [33, 47, 56, 65] and photometric stereo approaches [10, 60]. Since [10] allows for directional and point-light illumination, we consider both cases in our evaluation. [10]D and [10]P refers to the directional and point-light version, respectively. Given the calibrated nature of their framework, we furnish the estimated lighting parameters using our approach. Additionally, we explore a more challenging scenario wherein only two viewpoints with a wide baseline are available. We therefore introduce an ablated version of our framework, wherein the diffuse albedo is modeled with an MLP. This adaptation allows to highlight the advantages of employing the optimal diffuse albedo introduced in Sec. 5.2.

Full 3D reconstruction. As anticipated, [33, 47, 65] produce degenerate meshes in our wide baseline setup. Results in the supplementary demonstrate that they perform adequately when the distance between cameras is sufficiently small, although our framework consistently outperforms them. Conversely, Tab. 1 and Fig. 3 illustrate that our approach surpasses relevant baselines [10, 56, 60] both quantitatively and qualitatively. Our method yields sharper details without visible artifacts. Furthermore, error maps (shown in the supplementary material) reveal that the geometry obtained with [60] exhibits global distortions. Additionally, Fig. 4 demonstrates that both [56, 60] perform poorly when confronted with complex textures, while our results remain unaffected. This suggests that their pre-trained priors failed

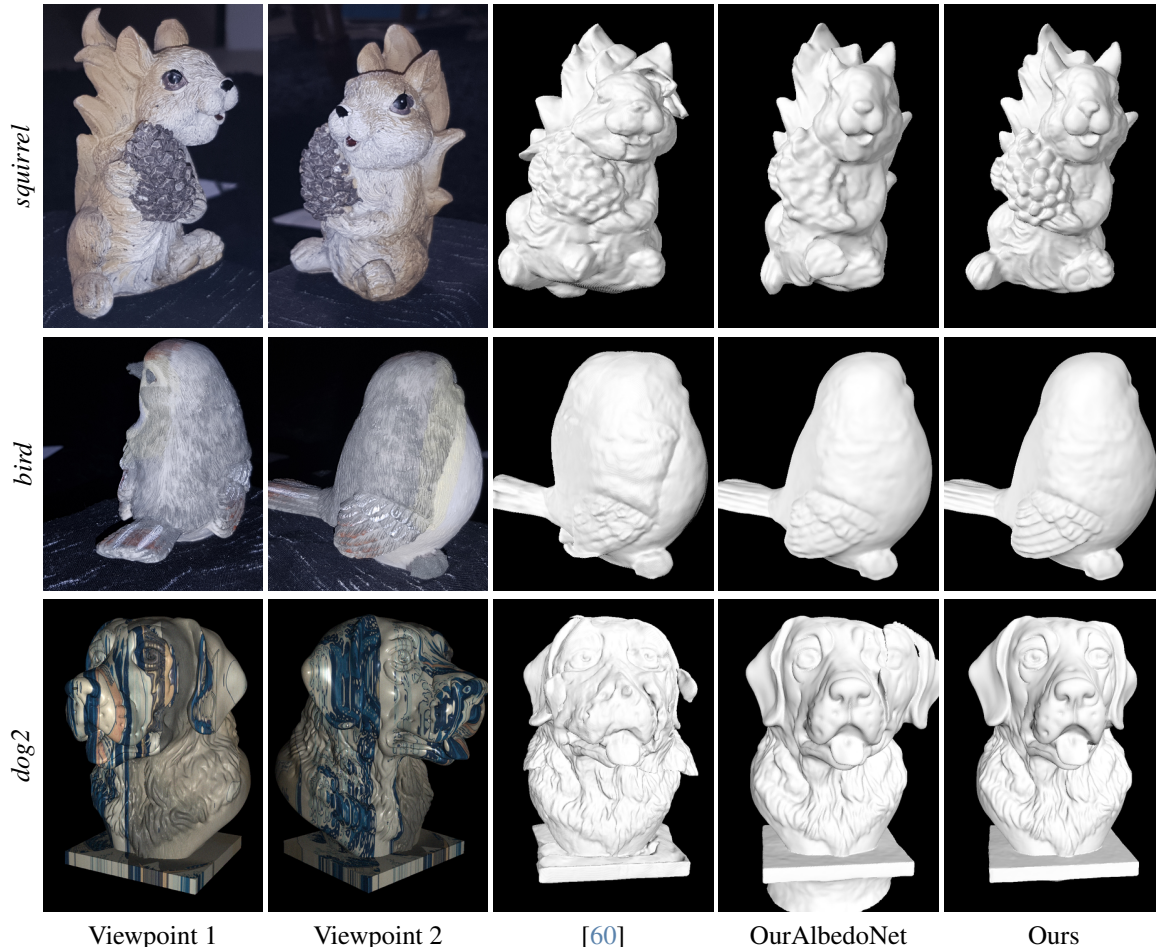| Viewpoint 1 | Viewpoint 2 | [60] | OurAlbedoNet | Ours |

Figure 5. Reconstructions from two distant viewpoints. Severe artefacts can be noticed for both [60] and *OurAlbedoNet*, and more particularly [60] introduce significant distortions in the shapes, as a result of using normal maps which do not provide strong enough constraints on the absolute position of the points.

to generalize for the given texture, possibly misassigning it as geometric patterns. This shows the advantage of approaches without relying on learned priors. The results for the DiLiGenT-MV dataset [28] are shown in the supplementary, demonstrating a high accuracy in the classical scenario of a dark room and distant light sources.

Two viewpoints. We denote our ablated framework, which uses a diffuse albedo network, as *OurAlbedoNet*. We exclude [56] from consideration as they require at least three viewpoints. Fig. 5 vividly illustrates the advantage of our framework in this challenging scenario, especially with the use of the optimal diffuse albedo. Notably, [60] and *OurAlbedoNet* exhibit pronounced artifacts and inaccuracies in geometric details in certain regions. Moreover, [60] results in severe distortions of the shape. We attribute this to their heavy reliance on per-view normal maps for reconstruction, which lacks absolute depth information about the shape. In contrast, our method directly estimates the shape from various point-light illuminations, coupled with a suitable model (Eq. (1)). This approach provides relevant cues

about the absolute position of the points, contributing to more accurate and high-quality reconstructions.

Limitations and future work. A dedicated section is available in the supplementary.

7. Conclusion

We introduced the first framework for multi-view uncalibrated point-light photometric stereo. It combines a state-of-the-art volume rendering approach with a physically realistic illumination model consisting of an ambient component and uncalibrated point-lights. As demonstrated in a variety of experiments, the proposed approach offers a practical paradigm to create highly accurate 3D reconstructions from sparse and distant viewpoints, even outside a controlled dark room environment.

References

[1] Meghna Asthana, William Smith, and Patrik Huber. Neural apparent brdf fields for multiview photometric stereo. In *Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10, 2022. 2, 3

[2] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 2

[3] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *European Conference on Computer Vision*, pages 294–311. Springer, 2020. 4

[4] Mohammed Brahimi, Yvain Quéau, Bjoern Haefner, and Daniel Cremers. *On the Well-Posedness of Uncalibrated Photometric Stereo Under General Lighting*, pages 147–176. Springer International Publishing, Cham, 2020. 2

[5] Mohammed Brahimi, Bjoern Haefner, Tarun Yenamandra, Bastian Goldluecke, and Daniel Cremers. Supervol: Super-resolution shape and reflectance estimation in inverse volume rendering. *arXiv preprint arXiv:2212.04968*, 2022. 2, 3, 5

[6] C Sidney Burrus. Iterative reweighted least squares. *OpenStax CNX. Available online: http://cnx.org/contents/92b90377-2b34-49e4-b26f-7fe572db78a1*, 12, 2012. 6

[7] Llukman Cerkezi and Paolo Favaro. Sparse 3d reconstruction via object-centric ray sampling. *arXiv preprint arXiv:2309.03008*, 2023. 2

[8] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8739–8747, 2019. 3

[9] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2

[10] Ziang Cheng, Hongdong Li, Richard Hartley, Yinqiang Zheng, and Imari Sato. Diffeomorphic neural surface parameterization for 3d and reflectance acquisition. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2, 3, 4, 6, 7

[11] Ziang Cheng, Junxuan Li, and Hongdong Li. Wildlight: In-the-wild inverse rendering with a flashlight. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2023. 2

[12] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 2

[13] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 2

[14] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2

[15] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[16] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2

[17] Heng Guo, Hiroaki Santo, Boxin Shi, and Yasuyuki Matsushita. Edge-preserving near-light photometric stereo with neural surfaces. *arXiv preprint arXiv:2207.04622*, 2022. 2, 4, 5

[18] B. Haefner, Z. Ye, M. Gao, T. Wu, Y. Quéau, and D. Cremers. Variational uncalibrated photometric stereo under general lighting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019. 2

[19] Bjoern Haefner, Songyou Peng, Alok Verma, Yvain Quéau, and Daniel Cremers. Photometric depth super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2453–2464, 2020. 2

[20] Bjoern Haefner, Simon Green, Alan Oursland, Daniel Andersen, Michael Goesele, Daniel Cremers, Richard Newcombe, and Thomas Whelan. Recovering real-world reflectance properties and shading from hdr imagery. In *2021 International Conference on 3D Vision (3DV)*, pages 1075–1084. IEEE, 2021. 4

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

[22] Shi-Sheng Huang, Zi-Xin Zou, Yi-Chi Zhang, and Hua Huang. Sc-neus: Consistent neural surface reconstruction from sparse and noisy views. *arXiv preprint arXiv:2307.05892*, 2023. 2

[23] Yakun Ju, Kin-Man Lam, Wuyuan Xie, Huiyu Zhou, Junyu Dong, and Boxin Shi. Deep learning methods for calibrated photometric stereo and beyond: A survey. *arXiv preprint arXiv:2212.08414*, 2022. 2

[24] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 3

[25] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. 3, 5

[26] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12601–12611, 2022. 2, 3

[27] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Multi-view photometric stereo revisited. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3126–3135, 2023. 2

[28] 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. 2, 3, 4, 7, 8

[29] Fotios Logothetis, Roberto Mecca, Yvain Quéau, and Roberto Cipolla. Near-field photometric stereo in ambient light. 2016. 2

[30] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. Semi-calibrated near field photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 941–950, 2017. 2

[31] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1052–1061, 2019. 2, 3

[32] Fotios Logothetis, Roberto Mecca, Ignas Budvytis, and Roberto Cipolla. A cnn based approach for the point-light photometric stereo problem. *International Journal of Computer Vision*, 131(1):101–120, 2023. 2

[33] 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*, pages 210–227. Springer, 2022. 2, 7

[34] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *Computer Graphics Forum*, pages 101–113. Wiley Online Library, 2021. 3, 4

[35] Roberto Mecca, Aaron Wetzler, Alfred M Bruckstein, and Ron Kimmel. Near field photometric stereo with point light sources. *SIAM Journal on Imaging Sciences*, 7(4):2732–2770, 2014. 2

[36] 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 *ECCV*, 2020. 2, 3, 4

[37] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H Kim. Practical svbrdf acquisition of 3d objects with unstructured flash photography. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018. 4

[38] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[39] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5569–5579. IEEE, 2021. 2, 3

[40] Thoma Papadhimitri and Paolo Favaro. Uncalibrated near-light photometric stereo. 2014. 2

[41] 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*, 39(8):1591–1604, 2016. 2, 3

[42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 4

[43] Rui Peng, Xiaodong Gu, Luyang Tang, Shihe Shen, Fanqi Yu, and Ronggang Wang. Gens: Generalizable neural surface reconstruction from multi-view images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[44] Songyou Peng, Bjoern Haefner, Yvain Queau, and Daniel Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 2

[45] Yvain Quéau, Tao Wu, and Daniel Cremers. Semi-calibrated near-light photometric stereo. In *Scale Space and Variational Methods in Computer Vision: 6th International Conference, SSVM 2017, Kolding, Denmark, June 4-8, 2017, Proceedings 6*, pages 656–668. Springer, 2017. 2

[46] 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*, 60:313–340, 2018. 2, 4

[47] Yufan Ren, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk, and Fangjinhua Wang. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16685–16695, 2023. 2, 4, 7

[48] Lu Sang, Bjoern Haefner, and Daniel Cremers. Inferring super-resolution depth from a moving light-source enhanced rgb-d sensor: A variational approach. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2020. 2

[49] Lu Sang, Björn Häfner, Xingxing Zuo, and Daniel Cremers. High-quality rgb-d reconstruction via multi-view uncalibrated photometric stereo and gradient-sdf. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3105–3114, 2023. 2

[50] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3716, 2016. 2

[51] Aditya Vora, Akshay Gadi Patil, and Hao Zhang. Divinet: 3d reconstruction from disparate views via neural template regularization. *arXiv preprint arXiv:2306.04699*, 2023. 2

[52] 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. *NeurIPS*, 2021. 2, 3, 4, 5

[53] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6187–6196, 2021. 2

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

[55] Chenglei Wu, Yebin Liu, Qionghai Dai, and Bennett Wilburn. Fusing multiview and photometric stereo for 3d reconstruction under uncalibrated illumination. *IEEE transactions on visualization and computer graphics*, 17(8):1082–1095, 2010. 2, 3

[56] Haoyu Wu, Alexandros Graikos, and Dimitris Samaras. S-volsdf: Sparse multi-view stereo regularization of neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3556–3568, 2023. 2, 3, 4, 6, 7, 8

[57] Luoyuan Xu, Tao Guan, Yuesong Wang, Wenkai Liu, Zhaojie Zeng, Junle Wang, and Wei Yang. C2f2neus: Cascade cost frustum fusion for high fidelity and generalizable neural surface reconstruction. *arXiv preprint arXiv:2306.10003*, 2023. 2

[58] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European conference on computer vision*, pages 674–689. Springer, 2020.

[59] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 2

[60] 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*, pages 266–284. Springer, 2022. 1, 2, 3, 6, 7, 8

[61] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2

[62] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 2

[63] 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*, 33:2492–2502, 2020. 2, 4

[64] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2, 3, 4

[65] Ying-Cong Chen Yixun Liang, Hao He. Rethinking rendering in generalizable neural surface reconstruction: A learning-based solution. *arXiv*, 2023. 2, 4, 7

[66] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 2

[67] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020. 2

[68] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6525–6534, 2021. 2

[69] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 3

[70] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 4

[71] Qing Zhang, Mao Ye, Ruigang Yang, Yasuyuki Matsushita, Bennett Wilburn, and Huimin Yu. Edge-preserving photometric stereo via depth fusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2479. IEEE, 2012. 4

[72] Dongxu Zhao, Daniel Lichy, Pierre-Nicolas Perrin, Jan-Michael Frahm, and Soumyadip Sengupta. Mvpsnet: Fast generalizable multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12525–12536, 2023. 2, 3