nLMVS-Net: Deep Non-Lambertian Multi-View Stereo

Kohei Yamashita Yuto Enyo Shohei Nobuhara Ko Nishino
Graduate School of Informatics, Kyoto University, Kyoto, Japan
https://vision.ist.i.kyoto-u.ac.jp/

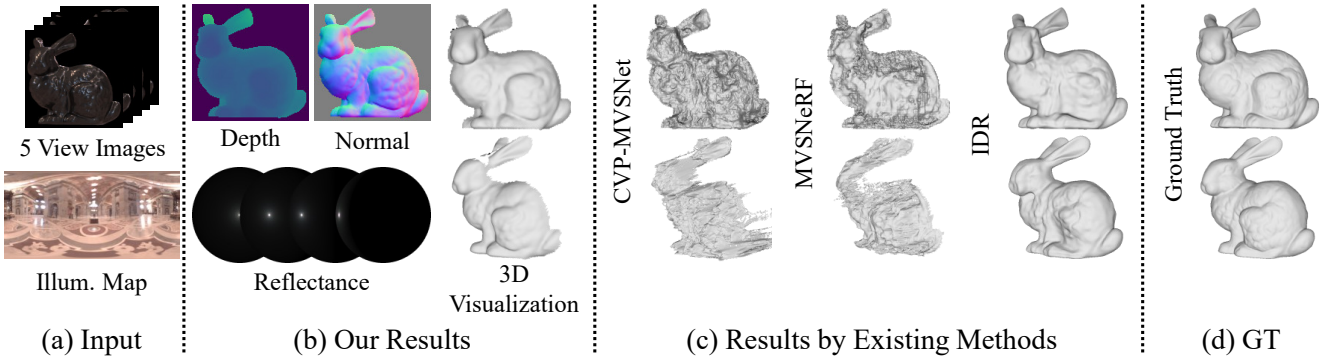


Figure 1: We introduce a novel MVS method that can jointly estimate per-pixel depths and surface normals together with the complex reflectance (b) of a textureless object from five images around the view of interest captured under known but natural illumination (a). The example results show that our method successfully recovers 3D shape consistent with ground truth (d) from a sparse set of images from which existing MVS and neural view synthesis methods struggle to recover accurate geometry (c). Please see, for example, the face and legs of the Stanford Bunny.

Abstract

We introduce a novel multi-view stereo (MVS) method that can simultaneously recover not just per-pixel depth but also surface normals, together with the reflectance of textureless, complex non-Lambertian surfaces captured under known but natural illumination. Our key idea is to formulate MVS as an end-to-end learnable network, which we refer to as nLMVS-Net, that seamlessly integrates radiometric cues to leverage surface normals as view-independent surface features for learned cost volume construction and filtering. It first estimates surface normals as pixel-wise probability densities for each view with a novel shape-from-shading network. These per-pixel surface normal densities and the input multi-view images are then input to a novel cost volume filtering network that learns to recover per-pixel depth and surface normal. The reflectance is also explicitly estimated by alternating with geometry reconstruction. Extensive quantitative evaluations on newly established synthetic and real-world datasets show that nLMVS-Net can robustly and accurately recover the shape and reflectance of complex objects in natural settings.

1. Introduction

Three-dimensional reconstruction of real-world objects of arbitrary reflectance is essential for many computer vision applications. In particular, a passive approach that can recover the 3D geometry for a view from a handful of images would be preferable. For downstream tasks such as scene navigation, object grasping, and augmented reality, explicit recovery of the reflectance properties in addition to the 3D geometry would be essential. As shown in Fig. 1(c), these requirements are hard to fulfill with neural view synthesis methods (*e.g.*, neural radiance field (NeRF)) as explicit geometry reconstruction is not their primary goal (*i.e.*, volume density only provides coarse view-dependent surface geometry) and as they usually require dense view sampling. Classic stereopsis and multi-view stereo (MVS) approaches [32], especially with recent integration of learned features and filtering, still excel in their simplicity, accuracy, and passive setup for explicit 3D geometry reconstruction.

Reconstruction of a textureless non-Lambertian surface (*e.g.*, a porcelain vase), however, still remains elusive to stereo-based approaches as stereopsis is limited by its two fundamental requirements: correspondence matching and triangulation. Finding correspondences directly translates

to making assumptions about the surface appearance, that they can be matched across views, *i.e.*, they are view-independent and texture-rich. This has largely limited the application of stereo-based methods to textured and Lambertian surfaces. Geometry recovery by triangulation also limits the output to surface depth which is often insufficient for capturing details of surface geometry.

Surface geometry can instead be recovered as part of inverting the radiometric process of image formation. Various methods have been proposed for single-view geometry reconstruction of non-Lambertian surfaces by jointly estimating its reflectance. The geometry recovered in such inverse rendering approaches is, however, fundamentally limited to surface normals. Although surface normals can expose finer surface details, they do not directly represent the surface.

In this paper, we introduce a novel multi-view stereo method that enables the simultaneous recovery of surface normals and depth for textureless non-Lambertian surfaces. At the same time, the method explicitly recovers the complex reflectance of the target surface. Our method is purely image-based. As we experimentally show in Sec. 4, it requires only a handful of (*i.e.*, 5) neighboring views for reconstruction from one vantage point. Most important, our method can be applied to objects with unknown complex reflectance captured under known but natural illumination. Our key idea is to integrate stereopsis with radiometric analysis so that radiometrically recovered geometric properties, namely surface normals, can serve as view-independent cues for multi-view stereopsis. We achieve this integration with an end-to-end learnable network which we refer to as nLMVS-Net.

Our nLMVS-Net consists of three key novel ideas. The first is a single-view shape-from-shading network that fully leverages radiometric likelihoods of surface normals. The network enables the estimation of per-pixel surface normal as a directional probability density which collectively serves as rich view-independent cues for subsequent multi-view stereo. The second key idea is a novel cost volume filtering network that leverages the recovered surface normal probability densities. The network integrates radiometric (*i.e.*, surface normals) and geometric (*i.e.*, correspondences) cues with a novel feature extraction layer and a consistency loss between the surface normal and depth estimates. The third is joint estimation of complex non-Lambertian reflectance by alternating with geometry estimation using a neural BRDF model [8].

We also introduce two newly collected datasets, which we refer to as nLMVS-Synth and nLMVS-Real. The synthetic dataset (nLMVS-Synth) consists of 26850 rendered images of 2685 objects with 94 and 2685 different real-world reflectance and natural illumination, respectively. We use nLMVS-Synth to train nLMVS-Net and thoroughly evaluate its accuracy on unseen synthetic images. The new multi-view image dataset of real objects, namely nLMVS-Real, consists of 2569 multi-view images of 5 objects each with one of 4 different reflectances taken under 6 different natural illuminations. Each of the 5 different objects is replicated with a 3D printer so that accurate ground truth geometry is available for quantitative analysis. This dataset is unprecedented in size for an accurately radiometrically and geometrically calibrated multi-view image set for a variety of surfaces and would undoubtedly serve as a useful platform for a wide range of shape reconstruction and inverse rendering research.

Experimental results on these datasets and others, together with direct comparisons with existing methods, clearly demonstrate the effectiveness of our method. We also show that the recovered depths and surface normals can be used to reconstruct a whole object from as few as 10 images. Thanks to the passive setup and sparse inputs, nLMVS-Net may prove useful for many 3D sensing applications including mobile sensing, XR immersion, and robotic navigation. All the data and code are publicly disseminated on our project web page.

	Depth	Normal	Reflectance Estimation	Sparse View Input	Nat. Illum.	General Object Category
MVSNet [38]	✓			✓	✓	✓
CVP-MVS [37]	✓			✓	✓	✓
NAS [22]	✓	✓		✓	✓	✓
RC-MVS [6]	✓			✓	✓	✓
Nam *et al.* [27]	✓	✓	✓			✓
Kaya *et al.* [18]	✓	✓				✓
Cheng *et al.* [9]	✓	✓	✓	✓		✓
Bi *et al.* [3]	✓	✓	✓	✓		✓
ON [29]	✓	✓	✓		✓	✓
NeRFactor [46]	✓	✓	✓		✓	✓
PhySG [45]	✓	✓	✓		✓	✓
IDR [41]	✓	✓			✓	✓
NeuS [34]	✓	✓			✓	✓
NeRS [44]	✓	✓	✓	✓	✓	
MVSNeRF [7]	✓			✓	✓	✓
Ours	✓	✓	✓	✓	✓	✓

Table 1: Image-based 3D reconstruction methods that exploit multi-view observations. Our method can recover geometry and reflectance from sparse (5 views) observations captured under known but natural illumination without any category-specific shape prior, which remains challenging to existing methods.

2. Related Work

We review relevant works on imaged-based 3D geometry reconstruction, mainly those methods that exploit multi-view observations. Table 1 summarizes the differences of our method and others. Our method can recover geometry and reflectance of textureless, non-Lambertian surfaces from a sparse set (*i.e.*, 5) of multi-view images captured under known but natural illumination without any category-specific shape prior, which remains challenging for existing

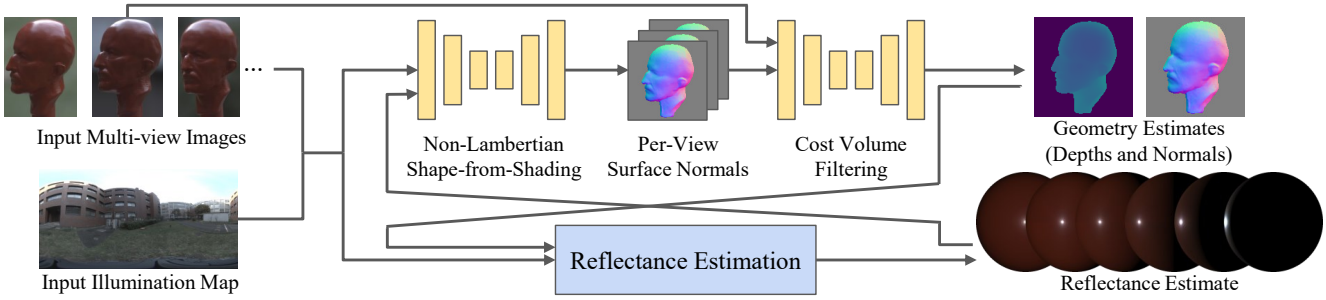


Figure 2: Overview of our novel multi-view stereo method, nLMVS-Net. The shape-from-shading sub-network learns to recover per-pixel probability densities of surface normals for each view. The novel cost volume sub-network then learns to reconstruct per-pixel depth and surface normals from those and the input sparse multi-view images. By alternating with neural reflectance estimation, we jointly recover the complex surface reflectance of the object.

methods.

Multi-view stereo relies on cross-view correspondence matching and triangulation. Traditional methods relied on manually designed distance metrics for correspondence detection and spatial aggregation [15, 11]. Recent works leverage deep neural networks to learn the metric directly from data. In particular, 3D convolutional neural networks are often used for cost volume filtering [38, 16, 39, 37]. Kusupati *et al*. [22] trained a deep neural network with a consistency loss to jointly estimate per-pixel depth and surface normal. The architecture of the cost volume filtering network of nLMVS-Net is inspired by this work, but fundamentally differs in that it handles not only textureless surfaces, but also non-Lambertian reflectance and even explicitly estimates it. Chang *et al*. [6] introduced a view synthesis loss that can implicitly handle non-Lambertian appearance. Their method, however, still relies on a photometric consistency loss which cannot handle large deviations from Lambertian reflectance.

Inverse rendering methods invert radiometric image formation to reconstruct object geometry [14, 17, 2, 13, 27]. Multi-view methods have exploited proxy geometry (*e.g*., a 3D mesh model) to recover surface geometry by iterative, nonlinear optimization [29, 35, 27, 45, 18, 19]. Oxholm and Nishino [29] alternated between updating a 3D mesh and BRDF parameters so that they are consistent with the input multi-view images and a known illumination map. These approaches struggle to recover high-frequency details of surface geometry as nonlinear optimization of geometry is unstable due to the large number of free parameters, especially when the number of input images is small (*e.g*., 20). A few methods handle sparse-view inputs [3, 9]. They, however, require collocated point lighting. In contrast, our method works with sparse inputs (5 images and an illumination map) under complex natural illumination.

Neural image synthesis methods recover a volumetric representation of a scene from a large number (typically on the order of tens to 100) of multi-view images [46, 45, 4, 41, 34]. Chen *et al*. [7] handled sparse (*i.e*., 3) views by conditioning the volume representation with input images. This method relies on volume rendering and the recovered volume density only provides view-dependent coarse depths. Zhang *et al*. [44] achieved explicit reconstruction of surface geometry and reflectance from sparse (*i.e*., 7) views by leveraging category-specific shape templates. They also recovered 3D shapes for arbitrary object categories by exploiting cuboids as templates. As we show in the supplementary material, this method struggles to generalize to objects with complex geometry that cannot be approximated with cuboids.

Multi-view stereo datasets have been proposed for benchmarking [1, 33, 21, 40]. They, however, are not radiometrically calibrated as they focus on Lambertian or textured surfaces rather than non-Lambertian objects. For the evaluation of non-Lambertian MVS, linear high dynamic range images that accurately capture the appearance of non-Lambertian objects are essential. Although the Multiview Objects under Natural Illumination database [28] provides high dynamic range images along with ground truth geometry and illumination maps, the number of instances is limited (4 objects under 3 environments). We also found that the images of this dataset contain flaws (please see Sec. 4). We introduce a novel real-world dataset that is accurate and extensive which can serve as a new platform for further studies on shape and reflectance recovery.

3. Deep Non-Lambertian Multi-View Stereo

Figure 2 depicts the overall structure of our model nLMVS-Net. The inputs are five multi-view images of an object and an illumination map of the surrounding environment. We assume the latter can be captured with a light probe, or can be estimated with a separate method [25, 12, 43]. Our nLMVS-Net consists of a single-view shape-from-shading network and a cost volume filtering

network. Since the appearance of non-Lambertian objects (*e.g.*, gloss) changes according to the viewing direction, we cannot directly achieve correspondence matching on the input images, especially for textureless surfaces. Instead, we explicitly extract view-independent features, namely surface normals, but while canonically accounting for uncertainty by encoding them as directional distributions with the shape-from-shading network. The recovered per-pixel surface normal distributions add rich information in addition to the regular appearance for multi-view correspondence matching and shape reconstruction. We derive a novel cost-volume filtering network that achieves seamless integration of these rich geometric and visual cues. We also derive a joint estimation framework that explicitly estimates the surface reflectance expressed by an invertible network together with the normals and depth.

3.1. Non-Lambertian Shape-from-Shading

The first step of nLMVS-Net is to recover the surface normals with associated uncertainties for each view. Surface normals naturally lie in a plausible range of directions for a given intensity as neither the illumination nor the reflectance is angularly unique [28]. As such, it is essential to model their uncertainties. For this, as depicted in Fig. 3a, we derive a novel deep neural network that estimates the per-pixel probability density distribution of surface normals for each view of the multi-view input images.

We assume an opaque, homogeneous reflectance for the object whose BRDF can be expressed as $\rho(\boldsymbol{\omega_i}, \boldsymbol{\omega_o}, \boldsymbol{n})$, where $\boldsymbol{\omega_i}$ is the incident direction, $\boldsymbol{\omega_o}$ is the viewing direction, and $\boldsymbol{n}$ is the surface normal. We also assume that the cameras and the illumination environment are distant from the object, *i.e.*, they can be approximated with an orthographic camera and an illumination map $L_i(\boldsymbol{\omega_o})$. Under these assumptions, the observed irradiance $E(\boldsymbol{\omega_o}, \boldsymbol{n})$ is

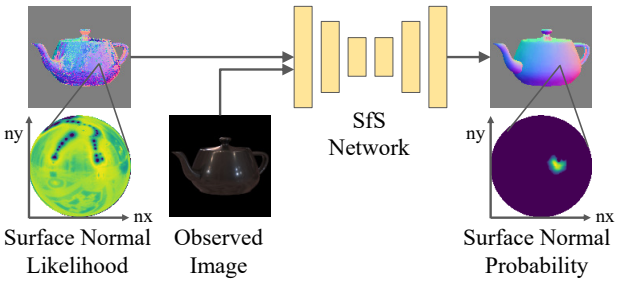
$$E(\boldsymbol{\omega_o}, \boldsymbol{n}) = \int L_i(\boldsymbol{\omega_i}) \rho(\boldsymbol{\omega_i}, \boldsymbol{\omega_o}, \boldsymbol{n}) \max(0, \boldsymbol{\omega_i} \cdot \boldsymbol{n}) \mathrm{d}\boldsymbol{\omega_i} \,. \tag{1}$$

We leave global light transport including shadows and interreflections for future work, and focus on object appearance by direct lighting which is dominant for single objects.

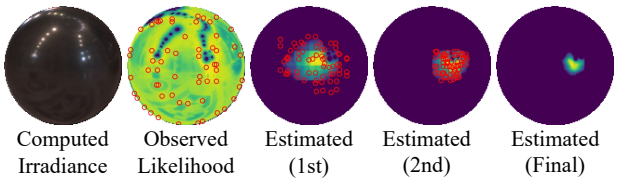
Let us assume that we are given a current estimate of the reflectance. This reflectance will be updated later in an alternating outer loop. For a given hypothesized surface normal and known illumination, its likelihood can be defined as the similarity of the irradiance $E(\boldsymbol{\omega_o}, \boldsymbol{n})$ computed from the given normal and the actual pixel value I:

$$p(I|\boldsymbol{n}) = \prod_k f(\log I^{(k)}; \log E^{(k)}(\boldsymbol{\omega_o}, \boldsymbol{n}), b) \,, \tag{2}$$

where $f(x; \mu, b)$ is the Laplace distribution and k is index of color channels. We use the Laplace distribution as its long tail is suitable for modeling deviations from the image formation model caused, for instance, by shadows and interreflections. We optimize the parameter b with training data. The surface normal directions are discretized with a 2D hemispherical grid and the likelihoods are computed for each direction.



(a) Shape-from-Shading (SfS) Network

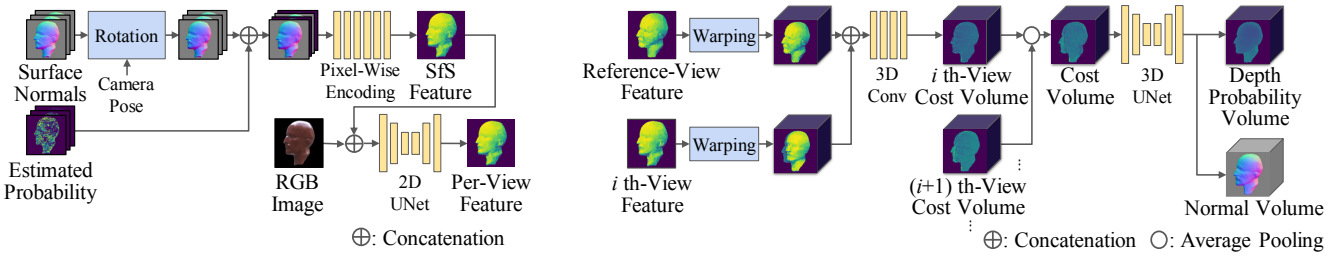


(b) Coarse-to-Fine SfS

Figure 3: (a) The shape-from-shading (SfS) network of nLMVS-Net learns to estimate pixel-wise probability densities of surface normals by aggregating local and global contextual information from the input view and observed pixel-wise likelihoods computed from the radiometric image formation model. (b) We use the SfS network recursively to refine the observed likelihoods in a coarse-to-fine manner. The red circles on the probability densities are the sampled surface normal orientations, which are used as inputs to the network in subsequent iterations.

The observed surface normal likelihoods Eq. (2) are too noisy and unreliable to use for cost volume filtering. We train the shape-from-shading network to refine and convert them into probability density distributions by aggregating local and global contextual information across the surface. As depicted in Fig. 3b, for computational efficiency, we achieve this in a coarse-to-fine manner. We first divide the possible surface normal orientations into a 8×8 grid and, for each grid, find the orientation that maximizes the observed likelihood $p\,(I|\boldsymbol{n})$ with brute-force search. We use the set of sampled surface normals and their observed likelihoods as inputs for each pixel. In the subsequent iterations, we double the resolution of the grid and sample surface normals around those that have high probability in the previous iteration. We use the same network with the same weights for all stages.

Since the inputs of the network (*i.e.*, sets of surface nor-



(a) Per-view feature extraction

(b) Cost volume filtering with 3D convolutional layers

Figure 4: The architecture of the cost volume filtering network of nLMVS-Net. A latent cost volume is constructed from the multi-view surface normal densities and color appearance which is then filtered with 3D convolutional layers. The network outputs two 3D volumes: depth probability volume and surface normal volume that encode the per-pixel depth and surface normal estimates as probability densities.

mals and their observed likelihoods) are unstructured, convolutions are not suitable for processing them. Instead, inspired by PointNet [30], we extract a 64 dimensional feature vector for each sampled surface normal and then fuse the sample-wise features by using max-pooling. The fused features are concatenated with features extracted from the input image and filtered by 2D convolutional layers. We use the pixel-wise filtered feature $\boldsymbol{a}$, a decoder MLP $g(\boldsymbol{n};\boldsymbol{a})$ that outputs a scalar value, and the observed likelihood distribution $p(I|\boldsymbol{n})$ to compute the output (unnormalized) probability density distribution $\hat{p}(\boldsymbol{n})$

$$\hat{p}(\boldsymbol{n}) = p(I|\boldsymbol{n})g(\boldsymbol{n};\boldsymbol{a})\,. \tag{3}$$

We train the network with images of synthetic objects whose BRDF and surface normals are known. In training, we compute the observed surface normal likelihoods by using the ground-truth BRDF and evaluate the network output with cross entropy loss

$$L_{\text{SfS}} = -\sum_i M(\boldsymbol{n_i}) \log\left(\frac{\hat{p}_i(\boldsymbol{n_i})}{\sum_j \hat{p}_j(\boldsymbol{n_j})}\right), \tag{4}$$

where $\{\boldsymbol{n_i}\}$ is the sampled (input) surface normal direction and M is a binary mask (1 iff the ground-truth and the sampled surface normal are in the same grid). The loss is evaluated for every stage of the coarse-to-fine estimation.

As shown in Figure 3a, the network significantly reduces the ambiguity of the observed likelihood distribution and extracts a well-defined probability density for each pixel.

3.2. Cost Volume Filtering

From the recovered per-pixel surface normal probability densities for each view as well as the original input images, nLMVS-Net learns to filter a cost volume to recover the object 3D shape as depth and surface normals. Figure 4 shows the architecture of the cost-volume filtering network.

As depicted in Fig. 4(a), the network takes in the multi-view input images as well as the outputs of the single-view shape-from-shading network, *i.e.*, per-pixel surface normal probability densities for each view. The latter are represented as sets of surface normals and their probabilities, from which we extract pixel-wise features. Since the surface normals are estimated in the camera coordinate system of each view, we first consolidate the coordinate system by rotating them using the known camera extrinsic parameters. We apply a feature extraction layer similar to the one in the shape-from-shading network (see Sec. 3.1) to convert the unstructured set of surface normals and their probabilities into a latent feature vector. The latent vector is concatenated with the input image and further filtered by a 2D UNet.

As illustrated in Fig. 4(b), the surface normal and image features are then used to construct a 3D latent cost volume which is then filtered with 3D convolutional layers. The final outputs are per-pixel depths and surface normals in the reference view. The outputs of the 3D convolutional layers become a depth probability volume $\hat{p}(d;\boldsymbol{m})$ and a surface normal volume $\hat{\boldsymbol{n}}(d,\boldsymbol{m})$ where d is a discretized hypothesis of depth. From these volumes, we compute the estimated depth $\hat{d}(\boldsymbol{m})$ and surface normal $\hat{\boldsymbol{n}}(\boldsymbol{m})$ as

$$\hat{d}(\boldsymbol{m}) = \sum_d \hat{p}(d;\boldsymbol{m})d\,, \tag{5}$$

$$\hat{\boldsymbol{n}}(\boldsymbol{m}) = \frac{\sum_d \hat{p}(d;\boldsymbol{m})\hat{\boldsymbol{n}}(d,\boldsymbol{m})}{\|\sum_d \hat{p}(d;\boldsymbol{m})\hat{\boldsymbol{n}}(d,\boldsymbol{m})\|}\,. \tag{6}$$

We ensure that the estimated depth and surface normals mostly agree with each other with a loss that aggregates the discrepancy in the directions of the surface normals and depth derivatives

$$L_{\text{dn}} = \sum_{\boldsymbol{m}} \arccos\left(\hat{\boldsymbol{n}}(\boldsymbol{m}) \cdot \bar{\boldsymbol{n}}(\boldsymbol{m})\right), \tag{7}$$

where $\bar{\boldsymbol{n}}(\boldsymbol{m})$ is the surface normal computed from the estimated depths $\hat{d}(\boldsymbol{m})$ with cross product of tangent vectors on the surface. As shown in Fig. 5, this consistency holds only in C0 and C1 continuous regions and we do not use $\bar{\boldsymbol{n}}(\boldsymbol{m})$

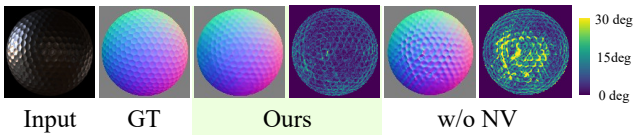


Figure 5: Advantage of estimating surface normals and depth as separate quantities. Our surface normal estimates are mostly consistent with the true surface normals (GT) in both continuous regions and also at surface discontinuities. In contrast, if depth derivatives (differentiation) are used as surface normals (*i.e.*, nLMVS-Net without normal volume–"w/o NV"), surface normals become erroneous at even continuous regions and high-frequency details of surface discontinuities are also lost or exaggerated.

as surface normal estimates. The summation in Eq. (7) ensures that strict consistency is only mildly imposed and surface normals are allowed to deviate from the depth derivatives at surface discontinuities. This is possible as we recover depths and surface normals as separate quantities.

We also impose individual depth and surface normal supervisions

$$L_{\mathrm{d}} = \sum_{\boldsymbol{m}} \|\hat{d}(\boldsymbol{m}) - d_{\mathrm{gt}}(\boldsymbol{m})\|_1 \,, \tag{8}$$

$$L_{\mathrm{n}} = \sum_{\boldsymbol{m}} \arccos\left(\hat{\boldsymbol{n}}\left(\boldsymbol{m}\right) \cdot \boldsymbol{n}_{\mathrm{gt}}(\boldsymbol{m})\right) \,, \tag{9}$$

where $d_{\mathrm{gt}}(\boldsymbol{m})$ and $\boldsymbol{n}_{\mathbf{gt}}(\boldsymbol{m})$ are the ground-truth depth and surface normal. The overall training loss is the weighted sum of these loss functions. We train this network separately from the shape-from-shading network.

3.3. Joint Shape and Reflectance Estimation

As Fig. 2 depicts, we alternate between estimating the object geometry and estimating the reflectance (BRDF). We represent the surface BRDF with the conditional invertible neural BRDF model (conditional iBRDF model) [8]. We update parameters of this model so that the difference between the input view images and the rendered images for each view is minimized. A challenge here is that pixel-wise intensity errors are too brittle as they suffer from geometry reconstruction errors. For this, we derive two objective functions that explicitly handle the reconstruction errors. The key ideas are that 1) we blur the images to evaluate the consistency in a coarse level, and that 2) we can find the ground truth surface normal around the estimated one that almost exactly satisfies the radiometric consistency and use it for rendering. Please see the supplementary material for further details.

3.4. Whole 3D shape Recovery

Our nLMVS-Net can recover per-pixel surface normal and depth of a reference view image from 5 input view images. If we have a set of such multi-view images that collectively cover the entire object (typically 10 images) taken while moving around an object on a plane, we can recover the whole 3D object geometry by applying nLMVS-Net to 5 images each while each image becomes the reference view, after which we integrate all depth and surface normals. During the alternating estimation of geometry and reflectance, we can use all the images together to update a single reflectance estimate. In the geometry estimation, for each view, we select four neighboring views (2 to the left and 2 to the right for a typical 360° capture on a plane) as inputs to nLMVS-Net. In the reflectance estimation, we compute the objective function introduced in Sec. 3.3 for each view and minimize the sum of them. We then reconstruct a 3D mesh from the estimated depth and surface normals by converting them into oriented points and applying Poisson surface reconstruction [20].

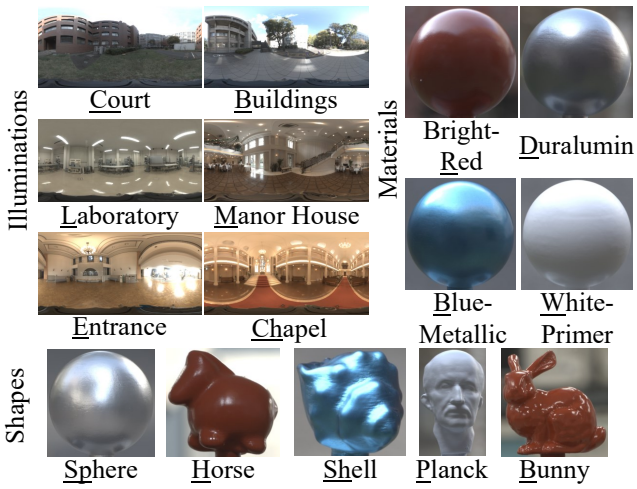


Figure 6: Our new multi-view real object image dataset, nLMVS-Real, consists of radiometrically and geometrically accurately calibrated 2569 multi-view images of 120 combinations of 5 shapes, 4 materials, and 6 illumination environments.

4. Experimental Results

We evaluate the effectiveness of nLMVS-Net through extensive experiments using both synthetic and real images of objects of different shapes and reflectances taken in a variety of illumination environments. For this, we introduce novel large-scale synthetic and real datasets, which we refer to as nLMVS-Synth and nLMVS-Real, respectively.

We first report that the Multiview Objects Under Natural Illumination Database [28] contains clear flaws in image capture including saturation, glare, and poor geometric calibration (please see the supplementary material for details). For this reason, numerical results on this dataset do not ac-

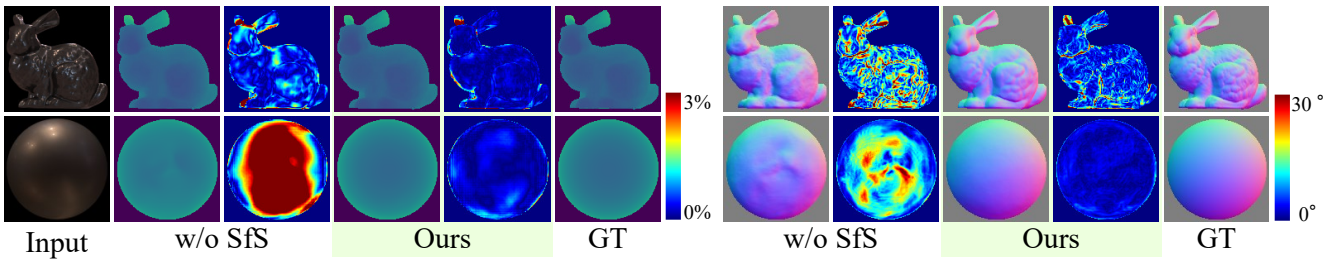


Figure 7: Ablation study on our shape-from-shading sub-network ("w/o SfS") which constructs a cost volume from multi-view image features without leveraging radiometric cues as per-pixel probabilistic surface normal likelihoods. For each result, we show the estimation errors as a color map. The results clearly show that the shape-from-shading sub-network is essential to handle textureless, non-Lambertian objects.

	Depth	Normal
RC-MVS [6]	5.76 %	-
MVSNeRF [7]	5.59 %	-
CVP-MVS [37]	4.16 %	-
IDR [41]	1.11%	11.0 deg
w/o SfS	1.12 %	11.2 deg
Ours	0.94 %	9.8 deg

(a) From 5 Views.

	Mesh
NeRS [44]	0.63 %
PhySG [45]	0.61 %
IDR [41]	0.25 %
Ours + PSR [20]	0.38 %

(b) From 10 Views.

Table 2: (a) Mean errors of the estimated depths and surface normals on the nLMVS-Synth dataset. The results clearly show the effectiveness of our method. (b) Mean error of 3D mesh models recovered from 10 view images. Even though we recover the mesh models by simply applying Poisson surface reconstruction (PSR) [20], our results are quantitatively comparable to the state-of-the-art methods.

curately reflect superiority of any method. This problem has been communicated with the authors of [28] and confirmed by them. In fact, one key contribution of our paper is the introduction of a new and a more extensive and accurate dataset that can replace this dataset. Radiometrically and geometrically accurate image capture for such dataset requires meticulous calibration and painstaking efforts in actual capture. Our new dataset (nLMVS-Real) would likely serve the community for a broad range of research on appearance modeling.

nLMVS-Synth Dataset We rendered a large number of training and test images of synthetic shapes [36, 24, 31, 5], measured BRDFs [26], and captured illumination maps [42, 12, 10, 23]. For training, we synthesized images of 2685 combinations of different shapes, materials, and illuminations. For each combination, we rendered images of randomly sampled 10 views. In total, the training set consists of 26,850 images. We also rendered a separate set of images for testing which consists of 4320 multi-view images of 216 different combinations of 6 shapes, 6 materials, and 6 illuminations. For each combination, we sampled 20 views on the horizontal line at equal intervals and added perturbations to them in the horizontal and vertical directions.

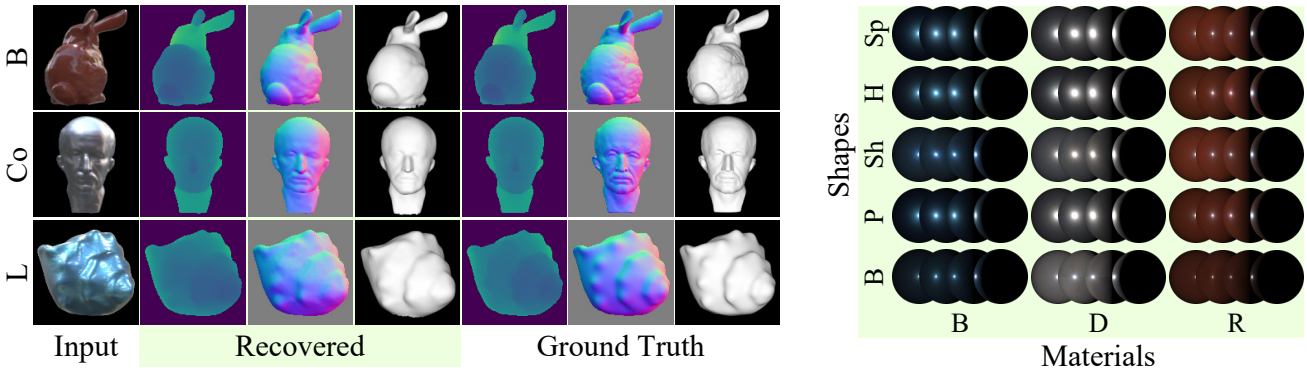
nLMVS-Real Dataset Figure 6 shows example images from our new nLMVS-Real dataset. We captured approximately 20 HDR images at three different heights for each of all combinations of 5 shapes, 4 materials, and 6 illumination environments. We also captured illumination maps using RICOH THETA Z1. The objects are replicated using a 3D printer and painted with different materials. Ground-truth 3D mesh models are available for quantitative evaluations.

Baseline Methods We compare our method with CVP-MVSNet [37], MVSNeRF [7], and RC-MVSNet [6] which also handle sparse (*i.e.*, 3 or 5 view) inputs. We could not directly compare our method with Kusupati *et al.* [22] as their network is trained to recover depths and normals for the entire scene rather than a single object. Instead, we compare our method with ours "w/o SfS" (without shape-from-shading) that constructs a cost volume only from multi-view image features similar to Kusupati *et al.* [22]. We also compare our method with IDR [41], a neural image synthesis method that recovers surface geometry from relatively sparse (*i.e.*, typically 50 and 11 at minimum) view inputs.

Evaluation Metrics We measure the accuracy of the recovered depth and surface normals with mean absolute errors. For depth error, the scale of the object is normalized such that the diagonal length of its bounding box is 1.

4.1. Results on Synthetic Data

Accuracy of the Shape-from-Shading Network We first evaluate the accuracy of the proposed shape-from-shading network. Ground-truth BRDF was used to compute the input surface normal likelihood; *i.e.*, we assume that the object's reflectance is known. The error between the ground-truth surface normals and those estimated to have the highest probability was lower than 10 degrees for 83% of all pixels. This is comparable to the accuracy of existing shape-from-shading methods such as Johnson and Adelson [17] and the single-view method of Oxholm and Nishino [29].



(a) (b)

Figure 8: (a) Recovered depths, surface normals, and whole 3D shape (mesh models) from our nLMVS-Real dataset. Please see supp. material for more results. Letters on the left denote different illumination environments. Our method successfully recovers accurate geometry for a wide variety of real-world objects. (b) Estimated BRDFs under "Court" illumination. Please see supp. material for all results. The results are consistent across estimation from images of different shapes, which demonstrates the accuracy of our method.

Joint Shape and Reflectance Estimation Results Figures 1 and 7, and Tab. 2a show qualitative and quantitative results. While existing methods and our method without the shape-from-shading cues (*i.e.*, "w/o SfS") fail on non-Lambertian and textureless objects, our method successfully recovers both depths and surface normal for these challenging objects. For 93 % of all input images, mean depth and surface normal errors were lower than 2 % and 19 degrees, respectively. These clearly show the effectiveness of our method. Please see the supplementary material for more results and ablation studies.

Accuracy of the Whole 3D Shape Recovery We can also recover 3D mesh models from our depth and surface normal estimates of 10 views (Sec. 3.4) and compare the results with those of neural image synthesis methods [44, 45, 41]. For this experiment, we used 10 views uniformly sampled from the original 20 views of the nLMVS-Synth dataset. We evaluate the reconstruction accuracy with root-mean-square (RMS) of the distance from a point on the reconstructed mesh to the nearest point on the ground truth mesh. Table 2b shows quantitative results. Even though we recover the mesh models by simply applying Poisson surface reconstruction [20], our results are quantitatively comparable to the state-of-the-art methods. Please see the supplementary material for qualitative results.

4.2. Results on Real Data

Figure 8a shows qualitative results of the recovered geometry on our nLMVS-Real Dataset. The results are of high quality even for complex shapes. Mean depth and surface normal errors were 2.01 % and 13.6 degrees, respectively. For 70 % of all input images, depth error was lower than 2 % and surface normal error was lower than 17 degrees. Figure 8b shows several of the estimated BRDFs. The estimates are consistent across different shapes. Note that ground truth BRDF of the real materials cannot be easily acquired.

As we make several assumptions about the objects and the capturing setup (*e.g.*, homogeneous material and distant illumination), the estimation accuracy would decrease for large deviations from these assumptions. Nevertheless, as the experimental results show, our method successfully recovers geometry and reflectance from real-world images that do not strictly satisfy them, which demonstrates the robustness of our method.

5. Conclusion

In this paper, we introduced nLMVS-Net, a neural multi-view stereo network that can recover both depth and surface normal at each pixel in the reference view for objects with complex reflectance taken under known but natural illumination. The method integrates radiometric cues in the form of view-independent surface normals recovered with a dedicated network into depth and surface normal cost volume filtering. By canonically modeling uncertainties of the surface normals, they provide rich cues for accurate geometry recovery. Experimental results clearly demonstrate the effectiveness of nLMVS-Net including its accuracy in recovering the complex reflectance of real-world objects. We believe nLMVS-Net can serve as a useful practical means for passive geometry recovery in the wild.

Acknowledgement This work was in part supported by JSPS 20H05951, 21H04893, JST JPMJCR20G7, JPMJSP2110, and RIKEN GRP. We also thank Shinsaku Hiura for his help in 3D printing.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-Scale Data for Multiple-View Stereopsis. *IJCV*, pages 1–16, 2016.

[2] Jonathan T. Barron and Jitendra Malik. Shape, Illumination, and Reflectance from Shading. *TPAMI*, 37(8):1670–1687, 2015.

[3] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3D Capture: Geometry and Reflectance From Sparse Multi-View Images. In *Proc. CVPR*, 2020.

[4] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition. In *Proc. NeurIPS*, 2021.

[5] John Burkardt. "PLY Files - an ASCII Polygon Format". https://people.sc.fsu.edu/~jburkardt/data/ply/ply.html.

[6] Di Chang, Aljaž Božič, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. RC-MVSNet: Unsupervised Multi-View Stereo with Neural Rendering. In *Proc. ECCV*, 2022.

[7] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. In *Proc. ICCV*, 2021.

[8] Zhe Chen, Shohei Nobuhara, and Ko Nishino. Invertible Neural BRDF for Object Inverse Rendering. In *Proc. ECCV*, 2020.

[9] Ziang Cheng, Hongdong Li, Yuta Asano, Yinqiang Zheng, and Imari Sato. Multi-View 3D Reconstruction of a Texture-Less Smooth Surface of Unknown Generic Reflectance. In *Proc. CVPR*, pages 16226–16235, 2021.

[10] Paul Debevec. "Light probe Image Gallery". https://www.pauldebevec.com/Probes/.

[11] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In *Proc. ICCV*, 2015.

[12] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to Predict Indoor Illumination from a Single Image. *ACM TOG*, 36(6), 2017.

[13] Aaron Hertzmann and Steven M Seitz. Example-Based Photometric Stereo: Shape Reconstruction with General, Varying BRDFs. *TPAMI*, 27(8):1254–1264, 2005.

[14] Berthold KP Horn. Shape From Shading: A Method for Obtaining the Shape of a Smooth Opaque Object From One View. Technical report, Massachusetts Institute of Technology, 1970.

[15] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast Cost-Volume Filtering for Visual Correspondence and Beyond. In *Proc. CVPR*, pages 3017–3024, 2011.

[16] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In-So Kweon. DPSNet: End-to-end Deep Plane Sweep Stereo. In *Proc. ICLR*, 2019.

[17] Micah K. Johnson and Edward H. Adelson. Shape Estimation in Natural Illumination. In *Proc. CVPR*, pages 2553–2560, 2011.

[18] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proc. CVPR*, 2022.

[19] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool. Neural radiance fields approach to deep multi-view photometric stereo. In *Proc. WACV*, pages 1965–1977, 2022.

[20] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

[21] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. In *Proc. SIGGRAPH*, 2017.

[22] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal Assisted Stereo Depth Estimation. In *Proc. CVPR*, 2020.

[23] ICT Vision & Graphics Lab. "High-Resolution Light Probe Image Gallery". https://vgl.ict.usc.edu/Data/HighResProbes/.

[24] The Stanford Computer Graphics Laboratory. "The Stanford 3D Scanning Repository". http://www.graphics.stanford.edu/data/3Dscanrep/.

[25] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality. In *Proc. CVPR*, June 2019.

[26] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A Data-Driven Reflectance Model. *ACM TOG*, 22(3):759–769, 2003.

[27] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H. Kim. Practical SVBRDF Acquisition of 3D Objects with Unstructured Flash Photography. In *Proc. SIGGRAPH Asia*, 2018.

[28] Geoffrey Oxholm and Ko Nishino. Multiview Shape and Reflectance from Natural Illumination. In *Proc. CVPR*, pages 2163–2170, 2014.

[29] Geoffrey Oxholm and Ko Nishino. Shape and Reflectance Estimation in the Wild. *TPAMI*, 38(2):376–389, 2015.

[30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. CVPR*, 2017.

[31] Szymon Rusinkiewicz, Doug DeCarlo, Adam Finkelstein, and Anothony Santella. "Suggestive Contour Gallery". https://gfx.cs.princeton.edu/proj/sugcon/models/.

[32] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002.

[33] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proc. CVPR*, 2017.

[34] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning Neural Im-

plicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Proc. NeurIPS*, 2021.

[35] Rui Xia, Yue Dong, Pieter Peers, and Xin Tong. Recovering Shape and Spatially-Varying Surface Reflectance Under Unknown Illumination. In *Proc. SIGGRAPH Asia*, 2016.

[36] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep Image-Based Relighting from Optimal Sparse Samples. In *Proc. SIGGRAPH*, 2018.

[37] Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu. Cost Volume Pyramid Based Depth Inference for Multi-View Stereo. In *Proc. CVPR*, 2020.

[38] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *Proc. ECCV*, 2018.

[39] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference. In *Proc. CVPR*, 2019.

[40] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks. In *Proc. CVPR*, 2020.

[41] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *Proc. NeurIPS*, 2020.

[42] Greg Zaal, Rob Tuytel, Rico Cilliers, James Ray Cock, Andreas Mischok, and Sergej Majboroda. "Poly Haven". https://polyhaven.com/hdris.

[43] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenmann, and Jean-François Lalonde. All-Weather Deep Outdoor Lighting Estimation. In *Proc. CVPR*, 2019.

[44] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In *Proc. NeurIPS*, 2021.

[45] 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 *Proc. CVPR*, 2021.

[46] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. In *Proc. SIGGRAPH Asia*, 2021.