A Unified Spatial-Angular Structured Light for Single-View Acquisition of Shape and Reflectance

Xianmin Xu[1]* Yuxin Lin[1]* Haoyang Zhou[1] Chong Zeng[1] Yaxin Yu[1] Kun Zhou[1,2] † Hongzhi Wu[1] †
[1]State Key Lab of CAD&CG, Zhejiang University [2] ZJU-FaceUnity Joint Lab of Intelligent Graphics

Abstract

*We propose a unified structured light, consisting of an LED array and an LCD mask, for high-quality acquisition of both shape and reflectance from a single view. For geometry, one LED projects a set of learned mask patterns to accurately encode **spatial** information; the decoded results from multiple LEDs are then aggregated to produce a final depth map. For appearance, learned light patterns are cast through a transparent mask to efficiently probe **angularly**-varying reflectance. Per-point BRDF parameters are differentiably optimized with respect to corresponding measurements, and stored in texture maps as the final reflectance. We establish a differentiable pipeline for the joint capture to automatically optimize both the mask and light patterns towards optimal acquisition quality. The effectiveness of our light is demonstrated with a wide variety of physical objects. Our results compare favorably with state-of-the-art techniques.*

1. Introduction

Joint acquisition of both shape and appearance of a static object is one key problem in computer vision and computer graphics. It is critical for various applications, such as cultural heritage, e-commerce and visual effects. Represented as a 3D mesh and a 6D Spatially-Varying Bidirectional Reflectance Distribution Function (SVBRDF), a digitized object can be rendered to reproduce the original look in the virtual world with high fidelity for different view and lighting conditions.

Active lighting is widely employed in high-quality acquisition. It probes the physical domain efficiently and obtains measurements strongly correlated with the target, leading to high signal-to-noise ratio (SNR) results. For geometry, structured illumination projects carefully designed pattern(s) into the space to distinguish rays for accurate 3D triangulation [15, 32]. For reflectance, illumination multiplexing programs the intensities of different lights over time, physically convolving with BRDF slices in the angular domain to produce clues for precise appearance deduction [13, 35].

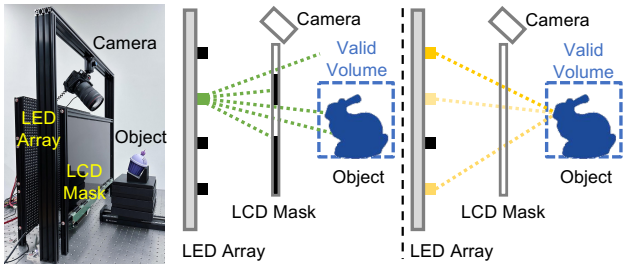

Figure 1. Our hardware prototype. It consists of an LED array, an LCD mask and a camera (left). One LED can project a set of mask patterns for shape acquisition (center), and multiple LEDs can be programmed to cast light patterns through a transparent mask for reflectance capture (right).

While active lighting for geometry or reflectance alone has been extensively studied, it is difficult to apply the idea to joint capture. At one hand, directly combining the two types of lights ends up with a bulky setup [39] and competing measurement coverages (i.e., a light for shape capture cannot be co-located with one for reflectance). On the other hand, existing work usually adopts one type of active light only, and has to perform passive acquisition on the other, leading to sub-optimal reconstructions. For example, Holroyd *et al*. [16] use projectors to capture geometry and impose strong priors on appearance. Kang *et al*. [19] build a light cube to densely sample reflectance in the angular domain. But the quality of its passive shape reconstruction is severely limited, if the object surface lacks prominent spatial features.

To tackle the above challenges, we propose a unified structured light for high-quality acquisition of 3D shape and reflectance. Our lightweight prototype consists of an LED array and an LCD mask, which essentially acts as a restricted lightfield projector to actively probe the spatial and angular domain. For geometry, each LED projects a set of mask patterns into the space to encode shape infor-

*Equal contributions.
†Corresponding authors (hwu@acm.org/kunzhou@acm.org).

mation. For appearance, the *same* LED array produces different light patterns, which are cast through a transparent mask to sample the reflectance that varies with the light angle. The prototype helps capture sufficient physical information to faithfully recover *per-pixel* depth and reflectance even from a single view.

To program the novel light towards optimal acquisition quality, we establish a differentiable pipeline for the joint capture, so that both mask and light patterns can be automatically optimized to harness our hardware capability. For geometry, a set of mask patterns are independently learned for each LED, by minimizing the depth uncertainty along an arbitrary camera ray. We also exploit the physical convolution that causes blurry mask projections in our setup, to encode richer spatial information for depth disambiguation. Multiple LEDs can be employed to project different sets of mask patterns, for improving the completeness and accuracy of the final shape. For reflectance, the light patterns are optimized as part of an autoencoder, which learns to capture the essence of appearance [19]. The reflectance is then optimized with respect to the measurements under such patterns, taking into account the reconstructed geometry for a higher-quality estimation.

The effectiveness of our approach is demonstrated on a number of physical samples with considerable variations in shape and reflectance. Using $4 \times 18 = 72$ mask patterns and 32 light patterns, we achieve on average a geometric accuracy of 0.27mm(mean distance) and a reflectance accuracy of 0.94(SSIM), on a lightweight prototype with an effective spatial resolution of only 320×320 and an angular resolution of 64×48. Our results are compared with state-of-the-art techniques on shape and reflectance capture, as well as validated against photographs. In addition, we evaluate the impact of different factors over the final results and discuss exciting future research directions.

2. Related Work

Due to space limit, here we mainly review geometry and/or reflectance acquisition techniques with *active illumination*. Interested readers are directed to excellent recent surveys for a broader view of the topic [7, 14, 31, 38, 40].

2.1. Shape Acquisition

This category of work can be divided into two groups, depending on whether the light samples the spatial or angular domain.

Highly accurate geometry can be captured with laser-stripe triangulation [23] or structured lighting [15, 26, 32]. These methods project single or multiple spatially distinctive patterns onto object surface, essentially encoding the light rays for subsequent 3D triangulation. Over the years, various patterns have been studied to improve robustness [15, 27], computational efficiency [9, 10] and acquisition speed [20]. Directly applying existing work to our setup results in a less satisfactory accuracy, due to the low spatial resolution and defocusing nature of our light. It is desirable to develop a pipeline that exploits our joint sampling capability/characteristics in the spatial and angular domain.

On the other hand, photometric stereo estimates a normal field from appearance variations with changing light angles. The result can be integrated into a depth map. Starting with the seminal work [41], substantial progress has been made to improve accuracy [4], efficiency [17] and robustness [2, 24, 33]. However, photometric stereo typically does not measure depths directly, and thus suffers from low-frequency shape distortions out of normal integration. In comparison, our light enables per-pixel measurements that are directly related to depths for accurate 3D reconstruction.

2.2. Reflectance Capture

Despite its high quality, exhaustive sampling a 6D SVBRDF on a known shape is prohibitively time consuming [6, 21]. One way to reduce the acquisition cost is to introduce priors over the reflectance [8, 22, 42] at the cost of compromised reconstruction quality. Another highly successful class of approaches are based on illumination multiplexing, which programs the intensities of lights at multiple angles, and recovers the reflectance from measurements under different lighting conditions. The lightstage reconstructs appearance from a pre-computed inverse lookup table [13]. Planar SVBRDF can be estimated from the appearance variation with respect to a moving linear light source [3, 12]. A frequency domain analysis is performed for capturing isotropic reflectance with an LCD panel as the light source [1].

Recently, differentiable acquisition techniques map both light patterns and the corresponding reconstruction algorithm to an autoencoder for an automatic, joint optimization. High-fidelity results are demonstrated on planar samples [18] and non-planar ones with structured [19] and unstructured conditions [25]. We build upon this line of work. One major difference is that we do not rely on network inference. Instead, our reflectance result is fine-tuned with respect to the measurements under optimized patterns. This leads to a higher reconstruction quality and more flexibility in handling challenging factors such as self-shadows with no compromise in acquisition efficiency.

2.3. Joint Estimation

Carefully engineered patterns in the angular domain are projected to sample appearance with an LED arc [35] or a light cube [19]. For each view, per-pixel reflectance maps are computed and then fed to a multi-view stereo algorithm. The geometry of textureless regions cannot be well recovered in passive shape reconstruction. Zhou *et al*. [44] capture different views of an object with circular LED lights.

Multi-view photometric stereo is applied to estimate the geometry, followed by reflectance computation. A gantry with a projector-camera pair is constructed in [16]. Phase-shift patterns are used to acquire depths, and a strong reflectance prior is imposed due to the sparse angular samples. Recently, Nam *et al*. [28] take hundreds of flash photographs from different views. The shape and reflectance are estimated with an involved alternating optimization.

Unlike our approach, none of the work above can efficiently probe *both* the spatial and angular domain, resulting in a tradeoff between geometry and reflectance reconstruction quality.

3. Hardware Prototype

Our lightweight acquisition setup consists of a spatial-angular light and a camera (Fig. 1). The light includes a rectangular RGB LED array and an LCD mask, which are parallel to and 15cm apart from each other. The LED array has $64 \times 48 = 3,072$ RGB LEDs, with a pitch of 1cm and a maximum total power of 240W. The intensity of each LED is independently controlled, and quantized with 8 bits per channel for FPGA implementation via pulse width modulation. The IPS LCD mask is ripped off from a conventional monitor, with a size of 59.8cm×33.6cm and a resolution of 1920×1080. It is directly controlled by on-device chip via HDMI. A 45MP Canon EOS R5 camera is mounted over the top edge of the mask, with a focal length of 24mm and an aperture of f/22. We define the valid volume of 3D points as a cube of 15cm×15cm×15cm, whose center is 15cm in front of the center of the mask. Physical objects are placed in this volume for acquisition.

Note that in mask pattern projection, we only use green LEDs to alleviate the undesirable spectral dispersion after passing the LCD. Due to the non-negligible size of each LED and the lack of dedicated optics for focusing light, the projected masks appear blurred on the object surface (Fig. 3). We use *binary* mask patterns, as two levels are sufficient for spatial encoding in our experiments and the potential tedious calibration of angular-varying transparency for every additional grayscale level can be avoided. Moreover, the effective spatial resolution of LCD mask is 320×320, as mask pixels not in this region will be projected to outside the valid volume.

Calibration. In addition to conventional camera intrinsic parameters/response curve calibrations and color correction, we calibrate the pose of each LED/mask as well as the spatial and angular LED intensity distribution: first, the LED of interest is turned on and the mask is set to a 2D array of 1x1 transparent squares with a spacing of 10 pixels; we then take photographs of the mask projection on a board with printed ARTags [11] that facilitate pose estimation; the board is placed at different poses to constrain the subsequent computation; finally, we minimize the differences between the photographs and the rendering results with the current estimates of parameters (Eq. 1), essentially performing differentiable calibrations. Please find more details in the supplemental material.

4. Overview

To capture a physical object from a single view, we first place it in the valid volume. For geometry acquisition, we set *learned* binary patterns to the LCD mask and turn on one of selected LEDs at a time, essentially projecting the patterns onto the object surface. Corresponding photographs are processed to produce a depth map result. Next, for reflectance capture, we set the LCD mask to fully transparent and program the intensities of the entire LED array according to *learned* light patterns. Finally, given the reconstructed shape, we perform differentiable optimization with respect to the image measurements to obtain the reflectance results, which are stored as spatially-varying BRDF parameters in texture maps. Please refer to Fig. 2 for an illustration of the pipeline and Fig. 3 for sample patterns and captured images.

5. Depth Acquisition

This section introduces depth acquisition with a single LED source, exploiting the characteristics of our setup. Below we first formulate our problem as code matching, then describe how to optimize mask patterns based on this formulation, and finally how to perform runtime computation. The extension to aggregate information acquired with different LEDs is introduced at the end.

5.1. Problem Formulation

Given a set of optimized mask patterns $\{M_j\}_j$, for each camera pixel visible to a particular LED, we collect a set of corresponding image measurements $\{I_j^g\}_j$ under such patterns. Our job is to take as input $\{I_j^g\}_j$ and output a depth along the camera ray of the current pixel. To do so, we first sample 3D candidates $\{\mathbf{x}_k\}_k$ along the line segment on the camera ray within the valid volume. For each $\mathbf{x}_k$, we simulate its image measurements $I_{j,k}^g$ under the mask pattern M_j via rendering with an area light source:

$$\begin{aligned} I_{j,k}^g &= \int_A L(\mathbf{x}_l, -\omega) M_j(\mathbf{x}_l \leftrightarrow \mathbf{x}_k) \rho F dA, \\ &\approx \rho F \int_A L(\mathbf{x}_l, -\omega) M_j(\mathbf{x}_l \leftrightarrow \mathbf{x}_k) dA, \\ &\approx \rho F L(-\omega) \int_A L(\mathbf{x}_l) M_j(\mathbf{x}_l \leftrightarrow \mathbf{x}_k) dA. \end{aligned} \quad (1)$$

Here $\mathbf{x}_l$ is a point on the current LED l, modeled as an area light of 2mm×2mm based on our calibration; $\mathbf{n}_l$/$\mathbf{n}_k$ is the surface normal of $\mathbf{x}_l$/$\mathbf{x}_k$, respectively. The light direction ω is a unit vector pointing from $\mathbf{x}_k$ towards $\mathbf{x}_l$. We

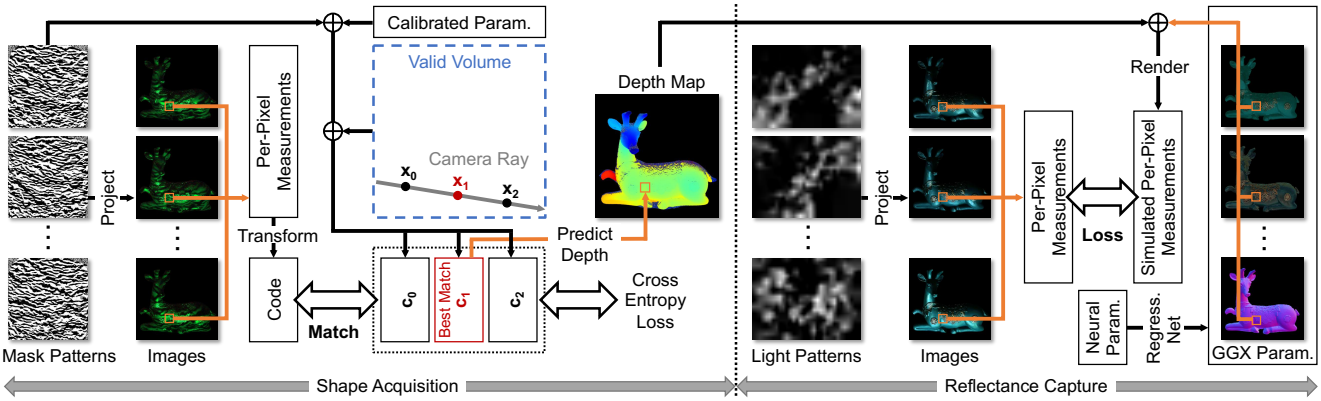

Figure 2. The complete pipeline. We first illuminate the physical project with learned mask patterns. For each valid pixel, we assemble the corresponding image measurements and transform into a code, which is matched against counterparts computed using sampled candidates along the current camera ray. The depth of the best matched result is selected as output. The masked patterns are pre-computed by optimizing a classification formulation among all codes of candidates. Next, we project pre-learned light patterns and also assemble the image measurements for each valid pixel. These measurements are compared with simulated measurements using a set of GGX parameters and the previously computed shape. Their difference is used to drive the optimization of a set of neural reflectance parameters, which produce the GGX parameters. The final reflectance results are stored as texture maps that represent different GGX parameters.

denote L as the emitted light from the LED, ρ the albedo, $F = \frac{(\omega \cdot \mathbf{n}_k)^+ (-\omega \cdot \mathbf{n}_l)^+}{||\mathbf{x}_l - \mathbf{x}_k||^2}$ the form factor, and $M_j(\mathbf{x}_l \leftrightarrow \mathbf{x}_k)$ the mask value where the ray from $\mathbf{x}_l$ to $\mathbf{x}_k$ intersects with the LCD. The integral is computed over the surface area A of l. Due to the small solid angle subtended by A with respect to $\mathbf{x}_k$, we assume constant $\rho/F/\omega$ across the integral and factor L as $L(\mathbf{x}_l, -\omega) = L(\mathbf{x}_l)L(-\omega)$ with $\int_A L(\mathbf{x}_l)dA = 1$. We implement $L(\mathbf{x}_l)$ as a 5×5 kernel, whose values are determined from calibration.

Next, we apply zero-mean and normalization [5] to $\{I_j^g\}_j$, to obtain a code $\mathbf{c}_k = [c_k^0, c_k^1, ...]$ that is independent of factors like albedo and form factor. Note that c_k^j is a scalar in the range of [0,1] and can be expressed as:

$$c_k^j = \alpha \int_A L(\mathbf{x}_l) M_j(\mathbf{x}_l \leftrightarrow \mathbf{x}_k) dA, \tag{2}$$

where α is a constant across different j. The above convolution encodes high-precision spatial information, as c_k^j *varies continuously* with changing $\mathbf{x}_k$. This is in contrast with an ideal point light that projects perfectly in-focused masks, as c_k^j would *stay the same* as long as $\mathbf{x}_k$ backprojects to the same mask pixel. Our superior sensitivity of c_k^j to $\mathbf{x}_k$(i.e., depth) leads to geometric reconstruction *beyond* the low spatial resolution of the LCD.

Now the problem can be formulated as follows. Given mask patterns $\{M_j\}_j$ and corresponding image measurements $\{I_j^g\}_j$, compute a code $\mathbf{c}$; match $\mathbf{c}$ with the codes simulated for all sampled candidates $\{\mathbf{x}_k\}_k$; output the depth of the best match $\mathbf{x}_{k_{\text{best}}}$.

5.2. Mask Pattern Training

Ideal mask patterns should make the codes $\{\mathbf{c}_k\}_k$ of candidates along a camera ray as distinguishable from each other as possible, to reduce mismatches that result in geometric errors. We cast this problem as standard multi-class classification, so that mask patterns can be optimized with a cross entropy loss.

Specifically, we randomly select a valid camera pixel first. Each pre-sampled candidate $\mathbf{x}_k$ along the corresponding camera ray is viewed as a different class. Next, we randomly pick $\mathbf{x}_t$ from $\{\mathbf{x}_k\}_k$ as the ground-truth class label. For each code $\mathbf{c}_k$, we compute its ZNCC score [5] as the dot product between $\mathbf{c}_k$ and $\mathbf{c}_t$. All scores then go through a softmax layer to produce a probability distribution, based on which a cross entropy loss is defined. The loss encourages a high probability for the labeled class. Now we can train $\{M_j\}_j$ for a given mask pattern number, since they are connected to the loss in a differentiable fashion (Fig. 2). Note that ZNCC is adopted here, because of its application in related work [26] to increase the robustness against factors including ambient illumination, varying albedos and acquisition noise.

In practice, each pixel in a mask pattern is initialized with a Gaussian noise. This value goes through a sigmoid function to fit in the valid range of [0,1]. Plus, we add a penalty that is increased over training iteration to encourage each mask pixel to be either 0 or 1. To increase robustness, we also add random zero-mean Gaussian noise to the synthetic measurements corresponding to $\mathbf{x}_t$ during training.

5.3. Runtime Computation

At runtime, for each valid camera pixel, we transform its image measurements $\{I_j^g\}_j$ to a code $\mathbf{c}$ via zero-mean and normalization. Next, the code is matched with its counterparts of sampled candidates $\{\mathbf{x}_k\}_k$ along the current camera ray. The candidate with the highest ZNCC score is selected to output a depth.

Below we extend our idea to multiple LEDs available in the setup. Geometric accuracy and completeness can be further improved, because the likelihood that a pixel is not visible to any LED often reduces and more LEDs impose more constraints to solve depth ambiguities [39]. Suppose we select n LEDs, each of which projects its own optimized mask patterns to the object, resulting in n sets of image measurements. For a valid camera pixel, we first determine its visibility with respect to each of n LEDs by applying [43]. It tests if any of the measurements, after subtracting their minimum, is above a certain threshold. Next, for each visible LED, we compute its code out of corresponding image measurements. Finally, these codes are concatenated into a single one, which is used to match its counterparts of the candidates $\{\mathbf{x}_k\}_k$ by compute ZNCC scores.

6. Reflectance Capture

With the reconstructed geometry, we compute the reflectance independently at each valid pixel via differentiable optimization. For a given pixel, our input is the image measurements $\{I_k^r\}_k$ at that location under a set of learned light patterns $\{L_k\}_k$. The output is the parameters for anisotropic GGX BRDF [37] (diffuse/specular albedos and roughnesses) as well as the shading frame (normal and tangent), which are stored in texture maps as the final result.

Light Pattern Training. While taking pictures with one LED on at a time is a straightforward way to sample appearance, it is highly inefficient due to the limited power of each LED and the sheer number of LEDs in the array. Therefore, we adopt [19] to learn a small number of light patterns that probe the reflectance in a compressive manner. Millions of lumitexels [22] are synthetically generated by the GGX BRDF model with random parameters along with a random local frame/position in the valid volume. These data, representing possible physical appearance, are used to train an autoencoder. Its encoder corresponds to light patterns used in acquisition, and the decoder is for computational reconstruction. Please refer to [19] for details.

Runtime Optimization. From image measurements under optimized light patterns $\{L_k\}_k$, state-of-the-art work [19] employs a decoder to predict lumitexels as output, which are subsequently fitted to GGX parameters. In this paper, we decide to keep the encoder only, as it captures key reflectance information efficiently; and we discard the original decoder and optimize the reflectance result by minimizing the differences between simulated and physical measurements under $\{L_k\}_k$ via differentiable rendering.

There are two reasons for developing our approach. First, compared with network inference in [19], our optimization better fits the measurements, producing appearance results that more closely match photographs. Next, it is easier to deal with challenging global effects. Specifically, self-shadows are not handled in [19], due to the difficulty in enumerating all possible visibility functions at training. In comparison, we can incorporate the reconstructed shape in differentiable rendering while simulating measurements (e.g., by computing visibility in the presence of the shape), resulting in a higher-quality estimation of reflectance.

Note that in differentiable rendering, we do not directly optimize GGX parameters. Instead, we reparameterize the GGX model with 16D neural parameters and jointly train 5 fully-connected networks, each of which transforms the neural parameters into one of the GGX parameters (Fig. 2) for each object. Please see the supplemental material for details. Compared with the original GGX model, our object-specific neural BRDF reparameterization is more amenable for deep learning and results in higher-quality reconstructions.

7. Results

We acquire the shape and reflectance of 7 physical objects from a single view. The maximum dimension of each object ranges from 9 to 15cm. A typical acquisition process uses $4 \times 18 = 72$ mask patterns and 32 light patterns. We set 20s of exposure time for each mask pattern, due to the limited power of a single LED, and 0.2s for each light pattern. Only LDR photographs are captured, with no HDR imaging. It takes about 24 minutes to finish the process. Subsequently, we manually specify a segmentation for each object to indicate regions of interest. In comparison with state-of-the-art work on geometric reconstruction, we use the projection of an object onto the camera as the segmentation result, regardless of the light visibility. PyTorch is used to implement the entire pipeline. We use Adam optimizer in all experiments. The computation is performed on a workstation with dual Intel Xeon 4210 CPUs, 256GB DDR4 memory and 4 GeForce RTX 3090 graphics cards. The reconstruction results are rendered with path tracing using OptiX.

Training. It takes 15 minutes to train 18 mask patterns for a single LED, using a learning rate of 5×10^{-4} and a batch size of 512. For each valid camera pixel, we sample candidates $\mathbf{x}_k$ with a density of 10 points/mm. For training 32 light patterns, it takes about 4 hours with the same learning rate of 5×10^{-4} and a batch size of 256.

Runtime. For geometry reconstruction, it takes 8 minutes to compute a depth map of about 1500×1500 with a batch size of 800, from measurements under a single LED. This time scales linearly with the number of LEDs

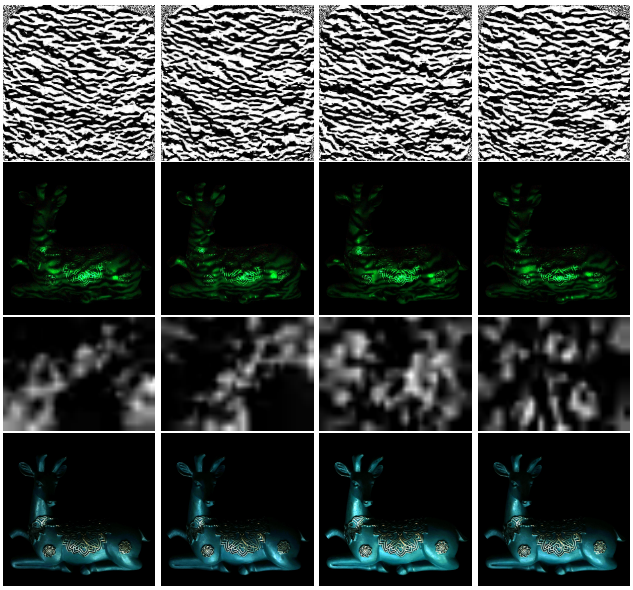

Figure 3. Visualization of learned mask patterns (1st row) and learned light patterns (3rd row). The corresponding photographs are listed one row below the patterns. Note that not all patterns/photographs are shown due to limited space.

used. To optimize the reflectance stored in texture maps of 1024×1024, it takes 1 hour with a learning rate of 1×10^{-3} and a batch size of 1024. The final appearance results are shown in Fig. 4.

Figure 4. Reflectance results, represented as GGX BRDF parameters. Note that tangent maps are not shown due to limited space.

7.1. Comparisons

Geometry. To facilitate quantitative analysis, we capture the "ground-truth" shapes of physical objects using a industrial handheld 3D scanner [34]. Geometric errors are reported in accuracy/completeness percentage at a 0.5mm threshold (denoted as A/C).

In the last three columns of Fig. 5, we compare against micro phase shifting (MPS) [15] and Gray code [36], which are two representative work on continuous/discrete shape encoding, respectively. Based on the effective spatial resolution of our LCD, a frequency-band of 16 pixels and 15 frequencies are used to generate 34 MPS patterns; in addition, 36 Gray code patterns, including the complement codes, are computed. Note that for robustness, we encode both x and y information with MPS and Gray code patterns. For a fair comparison, the same single LED is used to project different sets of mask patterns. Our results outperform existing techniques quantitatively. The periodic depth artifacts with [15] and [36] are due to their inferior performance with the low effective mask resolution and the blurred light projection in our setup (Eq. 1). For the grapefruit sample, we are not able to obtain a ground-truth shape due to its non-rigidity. Nevertheless, one can still visually compare the depth qualities.

Fig. 6 further compares with PS-FCN [4] and DVR [29], two state-of-the-art techniques on photometric stereo and single-image geometry estimation, respectively. For PS-FCN, we feed 192 images captured with a varying point source. Our results compare favorably against both methods, due to their lack of direct depth measurements or the mechanism for exploiting extra information beyond one single-view image.

Reflectance. We compare against two state-of-the-art methods [19] and [25], as well as validate against photographs taken at a novel lighting condition, in Figs. 5 and 7. In all cases, our results are of higher quality, due to the optimization that specifically fine-tunes with respect to the measurements of each object. Please refer to the accompanying video for animated results.

7.2. Ablations

We evaluate the impact of different factors over the results. In Fig. 8, the shape quality improves with the number of mask patterns. In Fig. 9 and the 4-5th columns in Fig. 5, both the 3D accuracy and completeness increase with the number of LEDs used in acquisition. These results show that our learning-based approach can automatically exploit the increase of input information for a better geometric reconstruction.

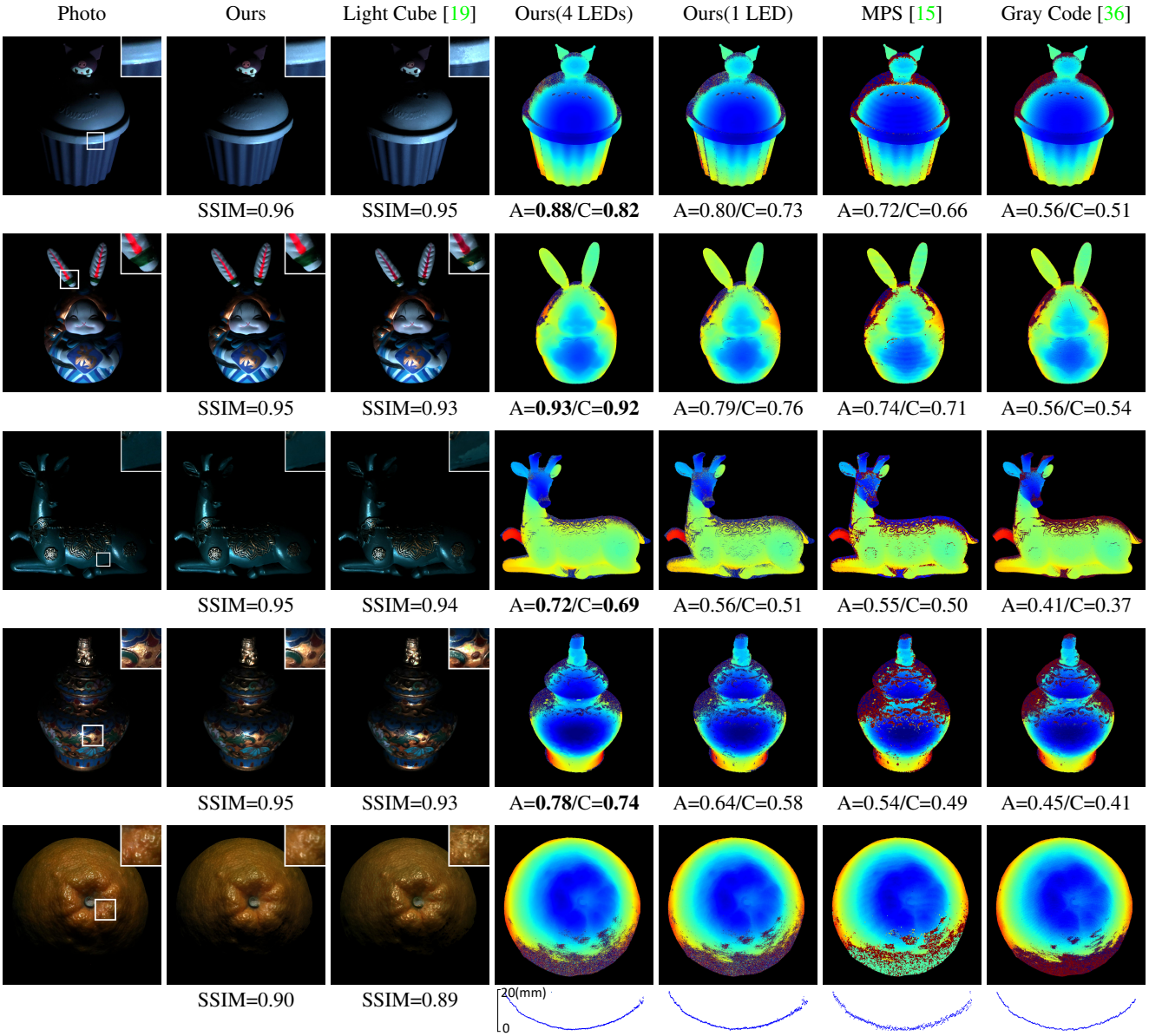

Figure 5. Comparisons with state-of-the-art techniques on shape and reflectance capture. From the left column to right, photograph with a lighting condition not used in our optimization, rendering with the reflectance results of our approach and light cube [19], depth maps reconstructed with our approach (4/1 LED), MPS [15] and Gray code [36]. Quantitative errors are reported in accuracy/completeness percentage and SSIM. Due to the difficulty in obtaining the ground-truth shape for grapefruit sample, 2D depth slices are shown instead.

Fig. 10 shows the results of using different mask patterns in conjunction with our pipeline. Our patterns outperform others, as the patterns are specifically optimized towards the goal of reducing depth error on this setup.

We study the impact of light patterns on reflectance quality in Fig. 11. First, our patterns outperform the same number of randomly point lights, which has a much smaller coverage of light directions. This demonstrates the superior angular sampling efficiency of our optimized patterns over point sampling. In the same figure, we test different numbers of light patterns. Our current choice of 32 strikes a good balance between reconstruction quality and acquisition speed, and is consistent with existing work [19].

8. Limitations and Future Work

Our work is subject to a number of limitations. First, for depth acquisition, a long exposure time must be set due to the drastically reduced radiances after going through the LCD even with a transparent mask. High-dynamic-range lightfield display with dedicated optics might be used to

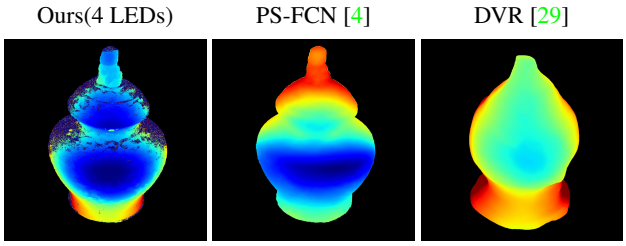

Figure 6. Comparison with state-of-the-art photometric stereo [4] and single-image shape estimation technique [29]. We use 192 input images for [4].

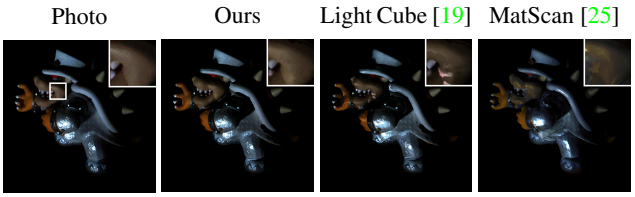

Figure 7. Comparison with state-of-the-art reflectance acquisition methods. From the left to right: photograph, our result, light cube [19] and using a handheld material scanner [25].

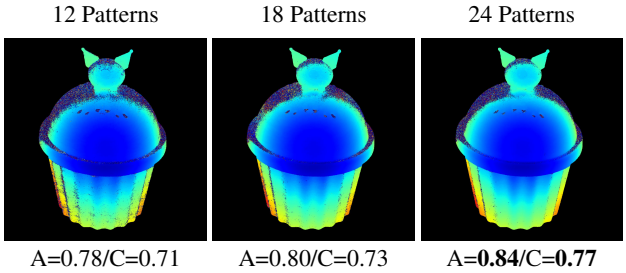

Figure 8. Impact of the number of mask patterns over reconstructed depths.

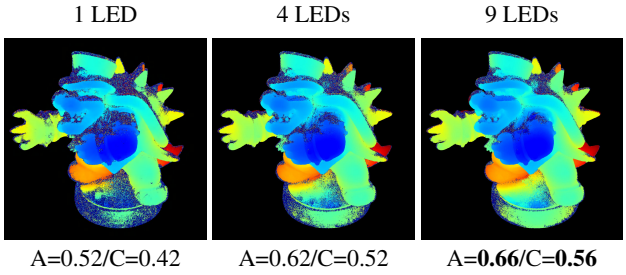

Figure 9. Impact of the number of LEDs used in acquisition over reconstructed depths.

substantially improve light transport efficiency, leading to a much shorter capture time. Next, the current set of multiple LEDs are manually selected and may not be optimal. It will be useful to automatically pick LEDs, based on a rough estimation of light visibility over the current object, for optimal reconstructions. Third, more complex light transport like interreflections are not considered in reflectance estimation, though applying a more advanced differentiable renderer [30] should solve this problem in a straightforward fashion. Moreover, we experiment with a single view only. A turntable can be added to scan a complete object from a sparse number of views.

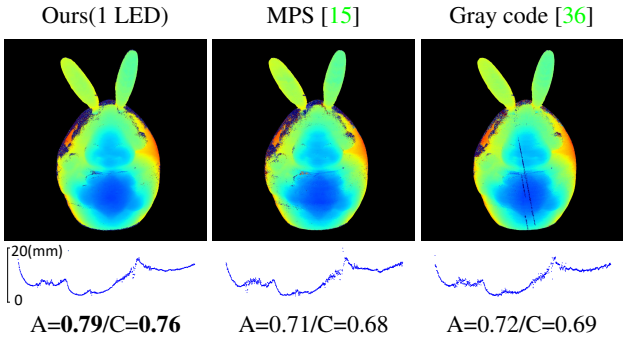

Figure 10. Impact of mask patterns over reconstructed depths. For the left to right, results from our optimized patterns, MPS [15] and Gray code [36] patterns used in conjunction with our pipeline.

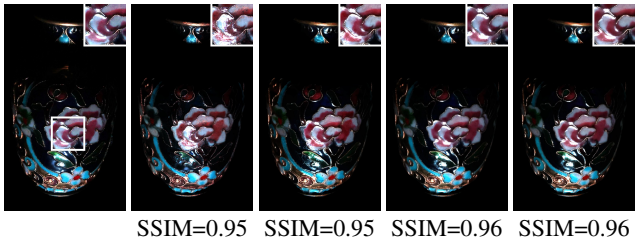

Figure 11. Impact of light patterns over recovered reflectance. From the left image to right, photograph, reflectance results from 32 randomly selected point patterns, 24, 32 and 48 optimized patterns.

We believe that this paper is only a small step towards structured illumination in the spatial-angular domain. It could open up many exciting research possibilities in the future. For example, instead of separately capturing shape and reflectance, it will be interesting to explore joint multiplexing of both LEDs and masks for improved acquisition efficiency. It is also promising to establish an adaptive pipeline for the joint capture. Last but not least, with specialized hardware, we are intrigued to develop a handheld scanner with spatial-angular structured illumination.

Acknowledgements. We would like to thank Xiaohe Ma, Kaizhang Kang, Weiwei Xu and Qi Sun for their help. This work is partially supported by NSF China (62022072 & 62227806), Zhejiang Provincial Key R&D Program (2022C01057) and the XPLORER PRIZE.

References

[1] Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. Practical SVBRDF capture in the frequency domain. *ACM Trans. Graph.*, 32(4):110:1–110:12, July 2013. 2

[2] Neil Alldrin, Todd Zickler, and David Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*, pages 1–8, 2008. 2

[3] Guojun Chen, Yue Dong, Pieter Peers, Jiawan Zhang, and Xin Tong. Reflectance scanning: Estimating shading frame and brdf with generalized linear light sources. *ACM Trans. Graph.*, 33(4):117:1–117:11, July 2014. 2

[4] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–18, 2018. 2, 6, 8

[5] J Crowley. Experimental comparison of correlation techniques. In *International Conference on Intelligent Autonomous Systems, 1995*, 1995. 4

[6] Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Trans. Graph.*, 18(1):1–34, Jan. 1999. 2

[7] Yue Dong. Deep appearance modeling: A survey. *Visual Informatics*, 2019. 2

[8] Yue Dong, Jiaping Wang, Xin Tong, John Snyder, Yanxiang Lan, Moshe Ben-Ezra, and Baining Guo. Manifold bootstrapping for svbrdf capture. *ACM Trans. Graph.*, 29(4):98:1–98:10, July 2010. 2

[9] Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, and Shahram Izadi. Hyperdepth: Learning depth from structured light without matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5441–5450, 2016. 2

[10] Sean Ryan Fanello, Julien Valentin, Christoph Rhemann, Adarsh Kowdle, Vladimir Tankovich, Philip Davidson, and Shahram Izadi. Ultrastereo: Efficient learning-based matching for active stereo systems. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6535–6544, 2017. 2

[11] Mark Fiala. Artag, a fiducial marker system using digital techniques. In *CVPR*, June 2005. 3

[12] Andrew Gardner, Chris Tchou, Tim Hawkins, and Paul Debevec. Linear light source reflectometry. *ACM Trans. Graph.*, 22(3):749–758, 2003. 2

[13] Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A. Wilson, and Paul Debevec. Estimating specular roughness and anisotropy from second order spherical gradient illumination. *Computer Graphics Forum*, 28(4):1161–1170, 2009. 1, 2

[14] Darya Guarnera, Giuseppe C. Guarnera, Abhijeet Ghosh, Cornelia Denk, and Mashhuda Glencross. Brdf representation and acquisition. *Computer Graphics Forum*, 35(2):625–650, 2016. 2

[15] Mohit Gupta and Shree K. Nayar. Micro phase shifting. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 813–820, 2012. 1, 2, 6, 7, 8

[16] Michael Holroyd, Jason Lawrence, and Todd Zickler. A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance. *ACM Trans. Graph.*, 29(4):99:1–99:12, July 2010. 1, 3

[17] Satoshi Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *ECCV*, 2018. 2

[18] Kaizhang Kang, Zimin Chen, Jiaping Wang, Kun Zhou, and Hongzhi Wu. Efficient reflectance capture using an autoencoder. *ACM Trans. Graph.*, 37(4):127:1–127:10, July 2018. 2

[19] Kaizhang Kang, Cihui Xie, Chengan He, Mingqi Yi, Minyi Gu, Zimin Chen, Kun Zhou, and Hongzhi Wu. Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Trans. Graph.*, 38(6):165:1–165:12, Nov. 2019. 1, 2, 5, 6, 7, 8

[20] Sanjeev J. Koppal, Shuntaro Yamazaki, and Srinivasa G. Narasimhan. Exploiting DLP illumination dithering for reconstruction and photography of high-speed scenes. *Int. J. Comput. Vis.*, 96(1):125–144, 2012. 2

[21] Jason Lawrence, Aner Ben-Artzi, Christopher DeCoro, Wojciech Matusik, Hanspeter Pfister, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Inverse shade trees for non-parametric material representation and editing. *ACM Trans. Graph.*, 25(3):735–745, July 2006. 2

[22] Hendrik P. A. Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Trans. Graph.*, 22(2):234–257, Apr. 2003. 2, 5

[23] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. The digital michelangelo project: 3d scanning of large statues. In *Proc. SIGGRAPH*, pages 131–144, 2000. 2

[24] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *CVPR*, pages 1490–1497, 2013. 2

[25] Xiaohe Ma, Kaizhang Kang, Ruisheng Zhu, Hongzhi Wu, and Kun Zhou. Free-form scanning of non-planar appearance with neural trace photography. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2, 6, 8

[26] Parsa Mirdehghan, Wenzheng Chen, and Kiriakos N Kutulakos. Optimal structured light a la carte. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2018. 2, 4

[27] Daniel Moreno, Kilho Son, and Gabriel Taubin. Embedded phase shifting: Robust phase shifting with embedded signals. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2301–2309, 2015. 2

[28] Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H Kim. Practical svbrdf acquisition of 3d objects with unstructured flash photography. In *SIGGRAPH Asia Technical Papers*, page 267, 2018. 3

[29] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 8

[30] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: a retargetable forward and inverse renderer. *ACM Trans. Graph.*, 38(6):203:1–203:17, 2019. 8

[31] Joaquim Salvi, Jordi Pagès, and Joan Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition*, 37(4):827–849, 2004. Agent Based Computer Vision. 2

[32] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003. 1, 2

[33] 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 *CVPR*, pages 3707–3716, 2016. 2

[34] Shining3D. EinScan Pro 2X Plus handheld industrial scanner. https://www.einscan.com/handheld-3d-scanner/2x-plus/. 6

[35] Borom Tunwattanapong, Graham Fyffe, Paul Graham, Jay Busch, Xueming Yu, Abhijeet Ghosh, and Paul Debevec. Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Trans. Graph.*, 32(4):109:1–109:12, July 2013. 1, 2

[36] Robert J Valkenburg and Alan M McIvor. Accurate 3d measurement using a structured light system. *Image and Vision Computing*, 16(2):99–110, 1998. 6, 7, 8

[37] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. Microfacet models for refraction through rough surfaces. In Jan Kautz and Sumanta N. Pattanaik, editors, *Proceedings of the Eurographics Symposium on Rendering Techniques, Grenoble, France, 2007*, pages 195–206. Eurographics Association, 2007. 5

[38] Michael Weinmann and Reinhard Klein. Advances in geometry and reflectance acquisition. In *SIGGRAPH Asia Courses*, pages 1:1–1:71, 2015. 2

[39] Michael Weinmann, Christopher Schwartz, Roland Ruiters, and Reinhard Klein. A multi-camera, multi-projector superresolution framework for structured light. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 397–404, 2011. 1, 5

[40] Tim Weyrich, Jason Lawrence, Hendrik P. A. Lensch, Szymon Rusinkiewicz, and Todd Zickler. Principles of appearance acquisition and representation. *Found. Trends. Comput. Graph. Vis.*, 4(2):75–191, 2009. 2

[41] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980. 2

[42] Hongzhi Wu, Zhaotian Wang, and Kun Zhou. Simultaneous localization and appearance estimation with a consumer rgb-d camera. *IEEE TVCG*, 22(8):2012–2023, Aug 2016. 2

[43] Yi Xu and Daniel G Aliaga. Robust pixel classification for 3d modeling with structured light. In *Proceedings of Graphics Interface 2007*, pages 233–240, 2007. 5

[44] Zhenglong Zhou, Zhe Wu, and Ping Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *CVPR*, pages 1482–1489, 2013. 2