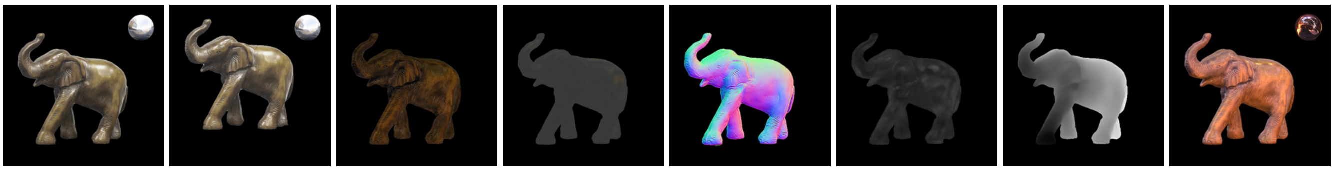
Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

Volume 42 (2023), Number 7

Deep Shape and SVBRDF Estimation using Smartphone Multi-lens Imaging

C. Fan[1,2] Y. Lin [2] A. Ghosh[1,2]

[1]Imperial College London, UK
[2]Lumirithmic Ltd.



(a) *Zoom lens* **(b)** *Wide-angle* **(c)** *Diffuse alb.* **(d)** *Specular alb.* **(e)** *Normal* **(f)** *Roughness* **(g)** *Depth map* **(h)** *Relighting*

Figure 1: *The estimated Bidirectional Reflectance Distribution Function (BRDF) and depth maps(c – g) of a wooden elephant are obtained from two input images (a, b) captured using a smartphone with a multi-lens imaging system. A probe is provided in the corners of input images to indicate the lighting conditions during the capture(not being used as an input). This enables realistic relighting rendering (h) under the Grace Cathedral lighting environment.*

Abstract

We present a deep neural network-based method that acquires high-quality shape and spatially varying reflectance of 3D objects using smartphone multi-lens imaging. Our method acquires two images simultaneously using a zoom lens and a wide angle lens of a smartphone under either natural illumination or phone flash conditions, effectively functioning like a single-shot method. Unlike traditional multi-view stereo methods which require sufficient differences in viewpoint and only estimate depth at a certain coarse scale, our method estimates fine-scale depth by utilising an optical-flow field extracted from subtle baseline and perspective due to different optics in the two images captured simultaneously. We further guide the SVBRDF estimation using the estimated depth, resulting in superior results compared to existing single-shot methods.

CCS Concepts
• ***Computing methodologies*** → ***Computational photography; Shape inference; Reflectance modeling;***

1. Introduction

Image-based shape and spatially varying reflectance estimation for 3D objects and materials have received significant research attention in the past few years. The approach has shifted from the traditional dense measurement methods towards making acquisition more practical, employing a minimal number of photographs while improving the acquisition quality. Recently, deep learning-based methods that require minimal acquisition from a single viewpoint have become popular due to their ease of use [LXR*18; BJK*20; DLG21].

In this work, we propose a deep neural network-based method that estimates high-quality shape and spatially varying reflectance of 3D objects using smartphone multi-lens imaging. Our method requires only two input images with subtle viewpoint and perspective shifts as captured by cameras on a smartphone. Modern multi-lens smartphones can acquire images with their two (or more) cameras simultaneously, making our method as effortless for acquisition as a single-shot method. Unlike previous work that utilises shading cues from flash illumination [LXR*18; BJK*20] or uses a combination of flash and polarization cues [DLG21], our method relies on the stereo information extracted from the two input images. This lifts the constraint on lighting, and hence the method works under either natural illumination or phone flash conditions. However, multi-lens on the back of a smartphone are usually only 2-3 centimetres apart, resulting in near sub-pixel disparity from corresponding features in two images. Acquiring an accurate depth map with such subtle view and perspective shifts and different optics is extremely challenging, and even state-of-the-art deep-learning-based methods for stereo matching [XZ20; LTD21; LWX*22] fail to delivery a reasonable result. Instead, we utilise RAFT [TD20] to estimate a fine-scale optical-flow field and use this optical-flow field as the shape cue for a network trained with a depth-based rendering loss for estimating the depth map for 3D objects. The depth-based rendering loss enables surface awareness for the network in training, producing more accurate, continuous and visually better results in inference. We further use the estimated depth to guide SVBRDF estimation, and we train a network as proposed in [DLG21]. We also demonstrate superior results in both shape and reflectance compared to other existing single-shot methods.

2. Related Work

We focus this overview on the most related work that aligns with the goals of this paper: shape and reflectance estimation under natural/flash lighting environments, using commodity hardware, and utilising deep learning. We also review the most recent deep stereo matching and optical-flow methods that are most relevant to this work.

2.1. Stereo Matching and Optical-flow

Dense stereo matching with deep neural networks has been an emerging research topic in computer vision, and various methods [PSR*17; KFR*18; YMHR19; GYY*19; THZ*20; XZ20; LTD21; LWX*22] have been proposed and each successive work shows improvements over it's predecessor on various stereo data-sets (e.g. KITTI [MHG15], ETH3D [SSG*17] and Middlebury [SHK*14]). One line of methods [PSR*17; THZ*20; XZ20] are 2D convolution based. Pang et al. [PSR*17] used multi-scale residual learning, AANet [THZ*20] proposed a novel aggregation method using sparse points and multi-scale interaction, [XZC*22], and [THZ*20] proposed a differentiable 2D geometric propagation and warping mechanisms to infer disparity. Another line of methods [KFR*18; YMHR19; GYY*19] calculates a cost volume between two input images, and filter the cost volume through 3D convolutions, which requires high computation and memory cost. Finally, a very recent method RAFT-Stereo [LTD21] adapted the iterative refinement in the optical flow network RAFT [TD20] to design a network for stereo matching. Similarly, Li et al. [LWX*22] proposed using Recurrent Update Modules with Adaptive Group Correlation that produces a much smaller cost volume than RAFT-Stereo[LTD21]. Last but not least, [XZC*22] uses a transformer network and formulates optical flow, rectified stereo matching and unrectified stereo depth estimation as a unified dense correspondence matching problem. However, all these methods aim at solving disparity for a scene with a significant camera baseline and the same optics, while we target estimating the object shape with a small baseline and differing optics as available on smartphone multi-lenses. Under similar small baseline conditions, Zhang et al [ZYR*22]. proposed MobiDepth, a real-time depth estimation system that effectively utilizes the dual cameras found on common mobile devices. This approach still utilizes the traditional SGBM (Semi-Global Block Matching) stereo-matching algorithm for depth detection, which still limits obtaining excellent results for distances greater than half a meter. Zhang et al's design is more aimed at detecting depth at a larger scale, such as in rooms or outdoors, rather than focusing on obtaining highly accurate depth maps within a small depth range. As the respective depth estimation problems differ fundamentally, we show these methods are not applicable to our data in section 3.2.

On the other hand, optical flow estimation methods [FDI*15; IMS*16; TD20] also establish dense pixel correlations between two input images, and estimate the per pixel movements in the image space. We demonstrate that the optical flow fields can act as cues for depth.

2.2. Practical Shape and Reflectance Acquisition

2.2.1. Commodity hardware

Recently there has been a focus on compact and portable capture methods employing commodity devices. Wu & Zhou [WZ15] have proposed an integrated system for hand-held acquisition of shape and reflectance of objects with a Kinect sensor. Aittala et al. [AWL15] have proposed a two-shot method for acquisition of stationary materials using a mobile phone. They employ a pair of flash-no flash observations of the sample coupled with statistical analysis to extract reflectance maps. The method has been extended to a single flash image for stationary materials using neural synthesis [AAL16]. Riviere et al. [RPG16] proposed two mobile acquisition setups for the acquisition of more general spatially varying planar surfaces. Free-form acquisition with flash illumination has also been employed for acquiring SVBRDFs of planar surfaces [HSL*17], and non-planar 3D objects [NLGK18]. Recent methods have demonstrated state-of-the-art results employing specialised LED panel/area-source [MKZ*21] or polarized sensor-flash pair [HJM*22] for such free-form acquisition.Xu et al. [XLZ*23] propose a method using structured light consisting of an LED array and an LCD mask coupled with a single-lens reflex camera to obtain shape and reflectivity from a single view. Wu et al. [WWZ16], Park et al. [PNS18] and Ha et al. [HBNK20] showed a series of progress in collecting geometry and SVBRDF, surface light field and shape in motion using RGB-D sensors.

However, the above family of approaches either rely on a large number of measurements and/or strong prior such as self-repetitive materials or planar geometry or use additional direct devices or sensors to get depth and SVBRDF.

2.2.2. Exploiting Deep Learning

Several methods have been proposed recently for surface reflectance and shape estimation from sparse measurements, including only a single observation by exploiting deep learning techniques. Many works have focused on SVBRDF estimation of planar samples under unknown environmental illumination [LDPT17; YLD*18], or uncalibrated flash illumination [DAD*18; LSC18] or both [DDB20], with further improvements using multiple flash measurements [DAD*19; GLD*19]. Deep learning has also been employed to estimate homogeneous reflectance properties of smooth convex objects of unknown shape under unknown illumination [GRR*17; MMZ*18]. Closer to our work, deep learning has been recently employed for joint shape and spatially varying reflectance estimation of non-planar objects from observations under flash illumination [LXR*18] or a combination of flash and ambient illumination [BJK*20]. Compared to these works, we demonstrate our approach to achieving higher-quality results by combining deep

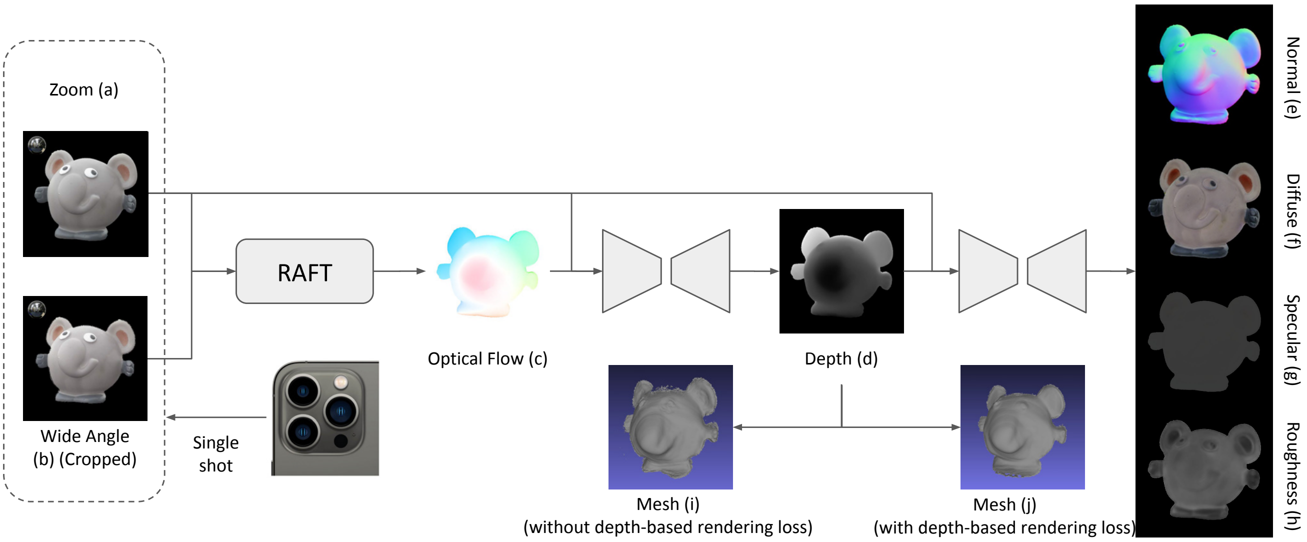


Figure 2: *Pipeline of our method for estimating shape and SVBRDF for real-world 3D objects. Our pipeline takes two images (a, b) respectively taken by the zoom and wide-angle lenses of a smartphone. We use a retrained RAFT [TD20] to estimate the optical flow (c). We then feed (a, c) to a depth estimation network to obtain the normalised depth (d), and forward (d) together with (a) to the final SVBRDF estimation network to get the remaining SVBRDF maps (e-h). With the proposed depth-based rendering loss, we can qualitatively tell the mesh (j) converted from the estimated depth is smoother and more accurate than the one in (i) obtained from the depth estimated using a network trained without the depth-based rendering loss.*

learning with multi-lens stereo cues, which allows our method also to work well under natural illumination. Also related to our approach are recent works that combine deep learning with polarization imaging for either shape estimation of 3D objects exhibiting homogeneous reflectance [BGW*20], or joint shape and spatially varying reflectance estimation under flash illumination [DLG21]. These methods either require a specialised polarization sensor or manual rotation of a polarization filter in front of a camera lens. Hence, while achieving high-quality results, these methods are more restricted compared to our proposed approach which employs single-shot like imaging using commodity mobile device and which can operate under both flash and natural illumination conditions.

3. Method

3.1. Overview

Our method aims at estimating both 3D shapes in terms of depth and normal, and spatially varying reflectance for real-world objects from two input images taken under either natural or flash lighting conditions, using smartphone multi-lenses. To tackle this highly ill-posed problem, we propose a two-stage method that first estimates the depth from a subtle camera baseline and perspective difference due to different focal lengths in two images acquired by the zoom lens and wide-angle lens respectively of a smartphone and use this estimated depth to guide the SVBRDF estimation in the second stage. Figure 2 demonstrates our two-stage pipeline.

Inspired by the recent stereo matching method [LTD21], we also leverage the RAFT network proposed by Teed et al. [TD20] and use it to establish dense pixel correlations and estimate the flow field which acts as the cue for our depth estimating network. We retrained the RAFT network using our data for the best quality result. Our depth estimation network uses the same architecture as proposed by Deschaintre et al. [DLG21], with a novel depth-based rendering loss. Unlike the depth rendering loss recently proposed by Chang et al. [CBZ*22] which uses a NERF style volumetric depth rendering, we solely utilize the predicted depth as the input for rendering and subsequently transform this predicted depth into surface normals. We designate specific parameters of the BRDF, such as roughness, to constant values in accordance with the conventional Cook-Torrance model and calculate the rendering results loss between the ground truth rendering. The incorporation of a depth-based rendering loss introduces a shading computation grounded in the depth gradient. This newly introduced loss term encourages greater surface continuity in the inferred object. As a result, the network's capacity to estimate smoother and more lifelike shapes for real-world 3D objects is improved.

Moving to the second stage, we feed the estimated depth to condition the SVBRDF estimation in the second network trained with our synthetic data. Again, we deploy a similar network as designed by Deschaintre et al. [DLG21]. We have experimented with different inputs to the network and found the network to work the best when there is perfect pixel correspondence between input images. Hence, we employ the zoom lens image as the main input image for both networks. We provide it together with the optical flow field as input for the depth prediction network in the first stage and we provide the zoom lens image together with the predicted depth map as input to the second network to predict the diffuse albedo, specular albedo, specular roughness and normal maps as output. To provide higher quality maps for subsequent rendering tasks, we performed enhancement on the obtained normal map in the end.

We employ an iPhone 13 Pro to acquire the two images for all our experiments. It is worth mentioning that the wide-angle lens image is cropped, and the zoom lens image is down-sampled to match the resolution of the cropped wide-angle lens image. We choose the zoom lens image as the main input image due to its higher resolution in raw as many details.

3.2. Depth Estimation

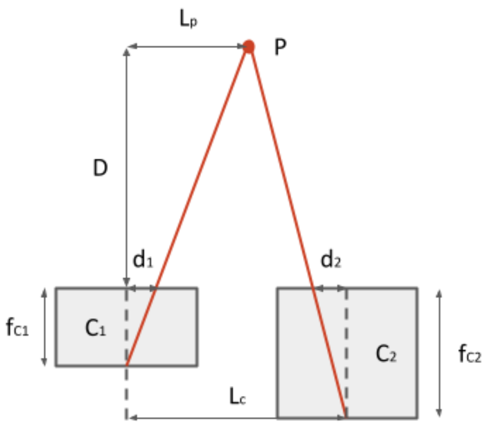


Figure 3: *A simple camera model: displacement of 3D points becomes ambiguous for the depth in disparity calculated from two cameras with different focal lengths.*

Given a camera C_1 with focal length f_{C_1} and another camera C_2 at L_C apart, with focal length f_{C_2}, and a point P at L_P away from the middle line of the camera C_1, with depth D to the imaging plane of both cameras (see Figure 3). The disparity of P between two imaging planes can be calculated as $\Delta dis = |d_1 - d_2|$, where

$$d_1 = \frac{L_P}{D + f_{C_1}} f_{C_1} \tag{1}$$

$$d_2 = \frac{L_P - L_C}{D + f_{C_2}} f_{C_2} \tag{2}$$

With $f_{C_1} = f_{C_2}$, it can be easily shown that Δdis does not depend on L_P, making the depth estimation trivial. However, in our case the two lenses have different focal lengths, resulting in an ambiguity between L_P and depth D. On the other hand, since L_C can be as short as 10mm, and L_P is quite narrow as we are solving for fine-scale depth of 3D objects, Δdis becomes sub-pixel which is beyond the precision of feature correspondence in existing stereo matching methods. Hence the depth is not resolved with even state-of-the-art deep learning based depth estimation methods [XZ20; LWX*22].

In the problem stated above, we are facing two challenges:

- Establish feature correspondence at sub-pixel or near sub-pixel level and calculate the disparity Δdis.
- Disambiguate the displacement L_P and depth D in the disparity Δdis.

We tackle the first feature correspondence challenge by utilising an optical-flow network, namely, RAFT [TD20]. With its dense 4D correlation volumes and recurrent optical flow updates, we find it generates the best off-the-shelf results compare to other optical flow networks [FDI*15; IMS*16]. Simultaneously, benefiting from fewer restrictions on lighting conditions by optical flow (only requires adjacent input images to be captured under the same lighting conditions) we win the flexibility on lighting conditions. To further improve the results we retrained the RAFT net with our data, details of which are described in section 3.5.

The trained optical flow net is an estimator of the ground truth optical flow f_o. We estimate the optical flow field that maps image I_1 to I_2, so the estimated optical flow $\hat{f_o}$ has per pixel correspondence to image I_1.

Next, to solve for the normalised depth, we train an estimator $\hat{F_D}$ that minimises $\mathcal{L} = ||\hat{F_D}(\hat{f_o}, I_1) - D||_1$. However, without surface awareness, the estimator $\hat{F_D}$ tends to jump locally while preserving the global shape. Instead, we redefine our loss:

$$\mathcal{L} = ||F_R(\nabla \hat{F_D}(\hat{f_o}, I_1)) - F_R(\nabla D)||_1 \tag{3}$$

where F_R is the rendering function, and $\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$. We find this loss enables local surface awareness for the network, and the network generates both numerically more accurate, and visually better quality results.

3.3. SVBRDF Estimation

Similar to the network architecture proposed in [DLG21], our SVBRDF estimation network jointly encodes the two inputs: the estimated depth $\hat{D}$ and the cropped image I_w taken by the smartphone zoom lens, and separately decodes through three decoders: a diffuse branch that outputs the diffuse albedo ρ_d, a specular branch that outputs the albedo ρ_s and roughness α_r, and finally a normal branch that decodes to the normal map normal. Our loss function as defined as:

$$\mathcal{L} = \mathcal{L}_d + \mathcal{L}_s + \mathcal{L}_\alpha + \mathcal{L}_{\vec{n}} + \lambda_R \mathcal{L}_R \tag{4}$$

where $\mathcal{L}_d$, $\mathcal{L}_s$, $\mathcal{L}_\alpha$ and $\mathcal{L}_{\vec{n}}$ are L_1 loss for diffuse albedo, specular albedo, roughness and normal respectiviely, and $\mathcal{L}_R$ is the rendering loss $||F_R(\rho_d, \rho_s, \rho_{\alpha_r}, \vec{n}) - F_R(\hat{\rho_d}, \hat{\rho_s}, \hat{\rho_{\alpha_r}}, \hat{\vec{n}})||_1$. we have chosen not to incorporate the depth-based rendering loss here as testing revealed that loss does not yield improvements.

3.4. Training

We modified RAFT[TD20] to output 32-bit vector files instead of the native 8-bit png images. We train separate networks for natural lighting and flash lighting with data specifically generated for each lighting condition respectively. Both networks were trained under 300,000 iterations to achieve convergence results. The training process took approximately 40 hours on a single A5000 card. The initial parameter settings, learning rate, and all other configurations follow conventional settings. The zoom lens image and wide-angle image were centred and cropped, and their resolution was reduced to 512*512 to serve as inputs for the network to manage the training memory efficiently. Details about training data generation are described in section 3.5.

3.5. Data Generation

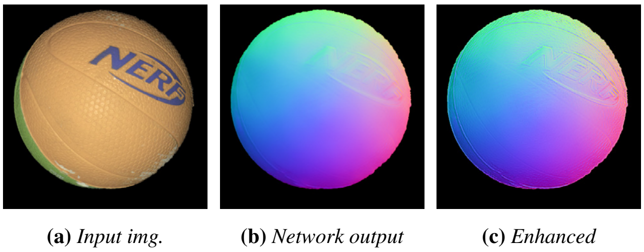
For training our networks, we require a large dataset of objects captured with smartphone multi-lenses under different lighting environments (including flash), along with ground truth labels for diffuse albedo, specular albedo, specular roughness, normal map, depth and optical flow field. Acquiring such a dataset would be extremely challenging and costly, hence we leverage synthetic data to create two datasets of over 150,000 and 5,000 sets of images respectively for natural lighting conditions and flash lighting. Our synthetic training dataset under natural lighting conditions is rendered using a combination of 50 complex meshes that well represent real-world 3D objects, 10 lighting environments that cover typical indoor and outdoor scenes, and 400 different materials (SVBRDFs). We further augment our lighting by rotating each environment map by $\pi/4$ and $-\pi/4$ along the x-axis, resulting in 30 environment maps. Our test dataset uses 10 meshes, 5 lighting environments and 30 materials not included in the training data. We replaced the lighting variance (30 EMs) with a frontal point light source for our flash lighting dataset. We fully simulate the perspective difference due to different focal lengths and small baseline as commonly seen on modern smartphone cameras. To better preserve fine details in both the optical flow field and depth map, we save the optical flow field as a 32-bit vector file and the depth map in a 16-bit PNG.

3.6. Surface Detail Enhancement

Due to the depth estimation being based on optical flow and the correspondence between the two views, our method may result in slightly blurred surface normals and attribute high-frequency surface detail to the diffuse albedo rather than the surface normal.

To address the limited surface detail in the normal maps, we employ a detail enhancement step similar to that proposed by [RPG16]. We apply **x** and **y** gradient filters to the estimated diffuse albedo. We then add these gradients with appropriate scaling to the **x** and **y** components of the estimated normal in the tangent space and renormalize the normal. Finally, we convert the modified normal back to world space. This process allows us to recover some high-frequency information in the normal map (see Fig. 4).

Although the enhanced details are not exact, they provide reasonable surface detail in the normal maps and improve the visual quality of the reconstructed surfaces for rendering purposes. We employ both the original predicted normal and the enhanced normal as diffuse and specular normal respectively for rendering using a hybrid normal rendering procedure similar to that proposed by [MHP*07].



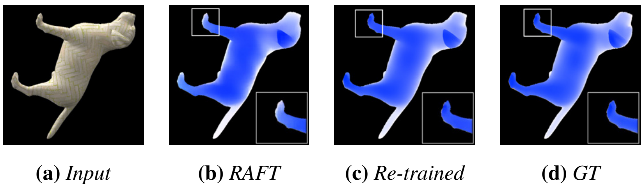
(a) *Input img.* **(b)** *Network output* **(c)** *Enhanced*

Figure 4: *An example of surface detail enhancement of normal map of a sponge ball. (b): the normal map directly get from network. (c): enhanced normal map.*

4. Result

4.1. Optical Flow

Figure 5 shows a qualitative comparison of the estimated optical flow between the native RAFT net and the retrained net using our data. After retraining, the estimated optical flow is of high quality, and it acts as the ambiguous shape cue for the depth estimation net to infer the detailed depth.



(a) *Input* **(b)** *RAFT* **(c)** *Re-trained* **(d)** *GT*

Figure 5: *Comparison of optical flow quality: 5b estimated optical flow from native RAFT [TD20]. 5c estimated optical flow from retrained RAFT using our data. The content in the white square is enlarged in the lower right corner to highlight the difference*

4.2. Depth and SVBRDF

We retrained three state-of-the-art stereo depth estimation networks [XZC*22; LTD21; LWX*22] using the same data as we used for training our network. Figure 6 shows comparison results. While these networks work great on datasets with large baselines and other extra constraints, e.g. x-axis consistency, they all fail to deliver good results for depth estimation in our specific setting.

In Figure 12, we demonstrate a few real-world 3D objects, where the first three rows are captured under a natural lighting condition, while the rest of the examples are captured with the phone flash (including one same test objects for comparison under natural vs flash illumination). Figure 1 is an additional example that has been acquired under natural illumination. In total, we present four sets captured under natural lighting conditions (2 indoors and 2 outdoors) and five sets captured with flash illumination. These examples demonstrate that our shape-from-optical-flow method ensures robust depth inference under different incident lighting conditions. The estimated depth information further regulates the estimation of the SVBRDF, leading to consistent quality results across the showcased examples.

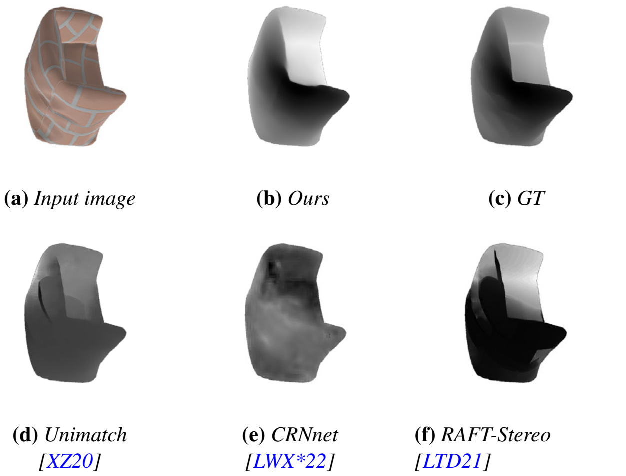


Figure 6: *Comparison of depth quality: depth estimated by existing methods 6d, 6e, 6f are nowhere close to GT in 6c, while ours 6b is accurate and detailed.*

Figure 7 shows some rendering comparisons of acquired objects to photographs under flash illumination from two different (novel) directions to further demonstrate the effectiveness of our method. Here, we selected the sponge ball shown in Figure 12(d) acquired under flash illumination, and the toy rubber elephant shown in Figure 2 acquired under natural indoor illumination for the validation against photographs under novel flash illumination conditions. It can be seen that our method correctly produces shading caused by the direction of light on the surface of the object, such as near the trunk of the toy elephant. We provide relighting videos of the acquired objects in the supplemental material as well.

We also illustrate our SVBRDF estimation results on synthetic data to directly compare to the ground truth. We do not apply the enhancement method to normal in this comparison. Figure 13 indicates our results preserve high-frequency details very well in both shape and reflectance, and renderings using our estimated maps are close to ground truth renderings.

4.3. Qualitative comparison

We compared our approach with previous successful methods in the next few sections. We chose the flash-based method of Li et al. [LXR*18] and the method of Boss et al. [BJK*20] relies on two photos - one with flash + ambient lighting and one with indoor ambient lighting only. To be clear, these two methods and our multi-lens methods are all single-view methods. Taking advantage of the multi-lens capabilities of smartphones, our method has an acquisition cost comparable to the single-input method of Li et al., and lower requirements than Boss et al.'s flash + no flash pair.

As can be seen from the comparison figure 8 on some synthetic results, our predicted reflectance and shape maps are qualitatively superior resulting in rendering that is much closer to the ground truth than these previous methods that mostly rely on flash illumination cues.

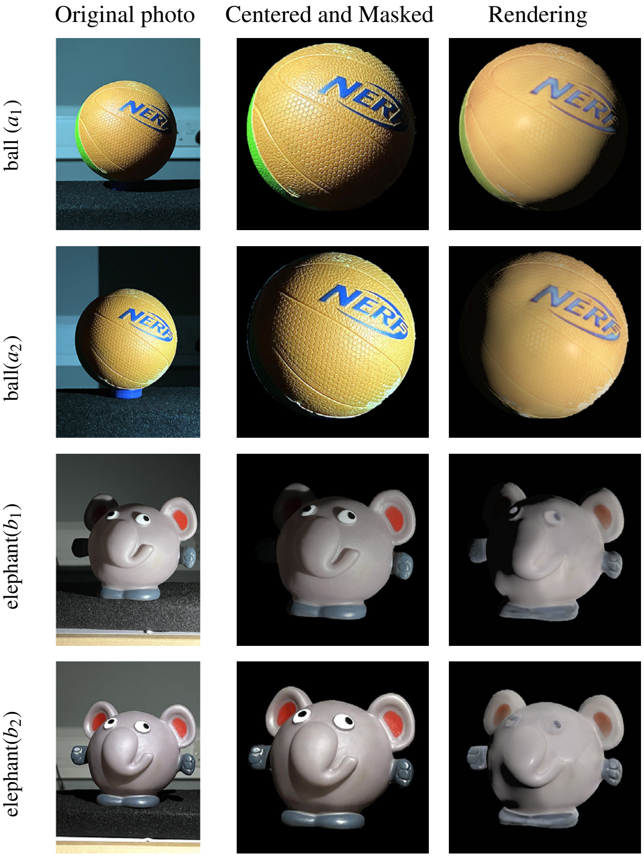


Figure 7: *Relighting comparisons of acquired objects under novel flash illumination conditions not employed for measurements. The sponge ball (a_1, a_2) has been acquired under frontal flash illumination. The toy elephant (b_1, b_2) has been acquired under natural illumination.The relighting comparison is under flash from 45 degrees top-left(a_1), 45 degrees bottom-right(a_2), right(b_1) and directly above (b_2).*

In Figure 9, we compare our results to the method of Li et al. [LXR*18] on two real objects captured with a phone flashlight. The results are consistent with our comparison and analysis of synthetic data which shows our method provides more accurate overall shape and BRDF estimation, while the method of Li et al. [LXR*18] suffers from significant low-frequency bias in the estimated depth and normals.

We also present a qualitative comparison to the method of Boss et al. [BJK*20] for an object acquired in outdoor conditions (Fig. 10). Boss et al.'s method strongly relies on flash illumination besides ambient illumination. However, due to the brighter outdoor lighting conditions, the flash illumination gets dominated by the ambient illumination. This results in very little difference between the two input images for their method and no strong flash cues which are required by their method. In comparison to results for the Rubik's Cube shown in Fig. 12(c) acquired in outdoor lighting with our method, the method of Boss et al. failed to provide reliable results in terms of normal and depth in Fig. 10. It is evi-

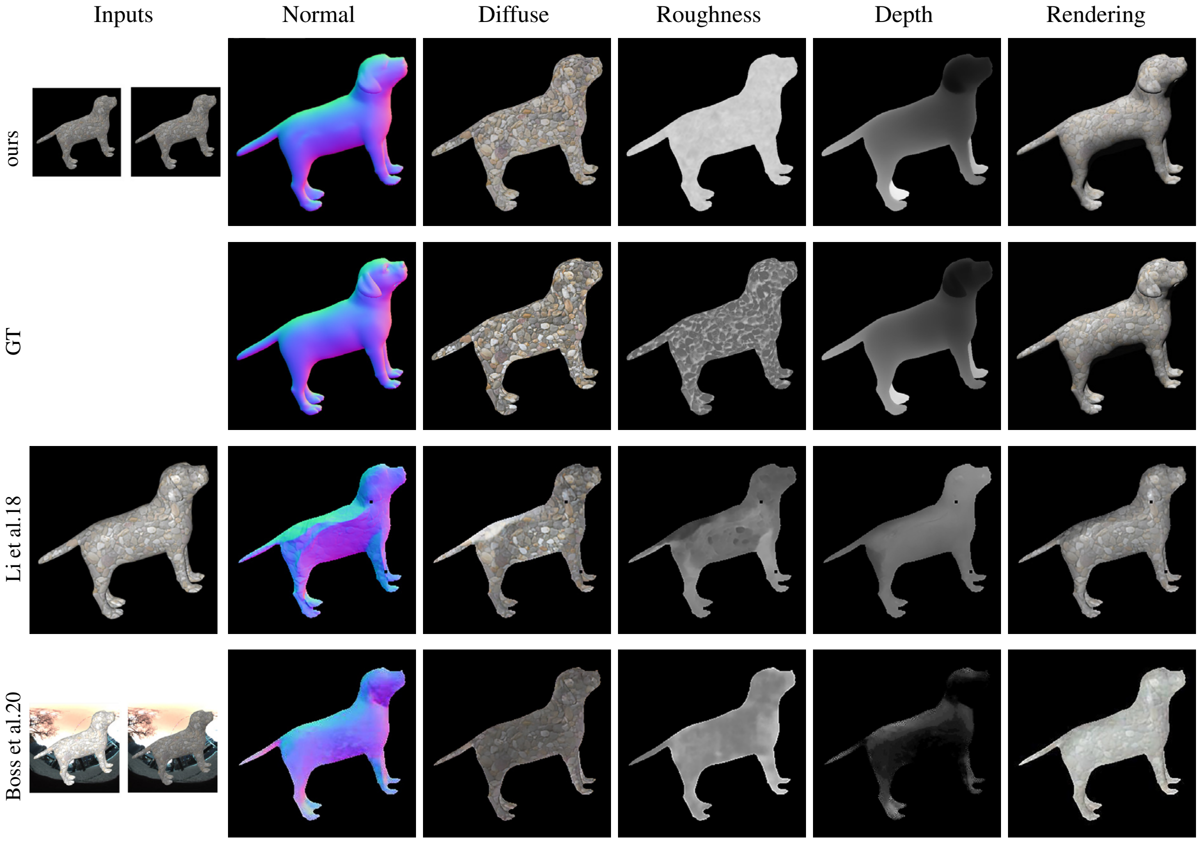
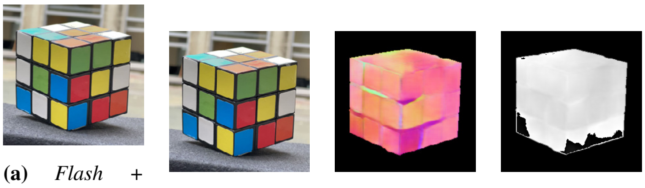


Figure 8: *We compare to the flash-based method of Li et al. [LXR*18] and the two-shots method of Boss et al. [BJK*20] on a synthetic example. Our estimated maps are closer to g.t. maps than results of [LXR*18; BJK*20], and our renderings are also a better match to the g.t.*

dent that such methods that strongly rely on flash illumination cues are not designed to handle scenarios involving strong ambient illumination, which limits their effectiveness in natural illumination conditions. In contrast, our method is able to handle both flash and natural illumination conditions due to depth estimated using multi-lens imaging.



(a) *Flash + amb.* **(b)** *Amb. only* **(c)** *Normal* **(d)** *Depth*

Figure 10: *A failure case for Boss et al.'s method [BJK*20] under strong natural illumination. Under strong natural lighting, it becomes difficult to distinguish between input images captured with (a), and without (b) flash. The absence of flash cues in the input results in large errors in estimated surface normal and depth.*

	Li et al.	**Boss et al.**	**Ours (Flash)**	**Ours (EM)**
Depth	0.2964	0.2763	0.0473	0.0350
Normal	0.1667	0.1637	0.0952	0.0888
Renderings	0.1252	0.2092	0.0610	0.0690

Table 1: *Quantitative comparison results. Numerically we perform significantly better compared to Li et al. [LXR*18] and Boss et al. [BJK*20] on both flash data and natural illumination data.*

4.4. Quantitative Comparison

In table 1 we compare quantitatively to Li et al. [LXR*18] and Boss et al. [BJK*20] on synthetic data using L1 difference. We evaluate the error on the normal maps, depth and renderings instead of SVBRDF maps, as different BRDF models have been chosen by the different methods. This quantitative evaluation is performed on 250 sets of synthetic input images, consisting of renderings produced using 5 meshes, with each randomly rotated 5 times, and 10 SVBRDF. The rendering error is computed over 20 renderings for each result with varying light properties.

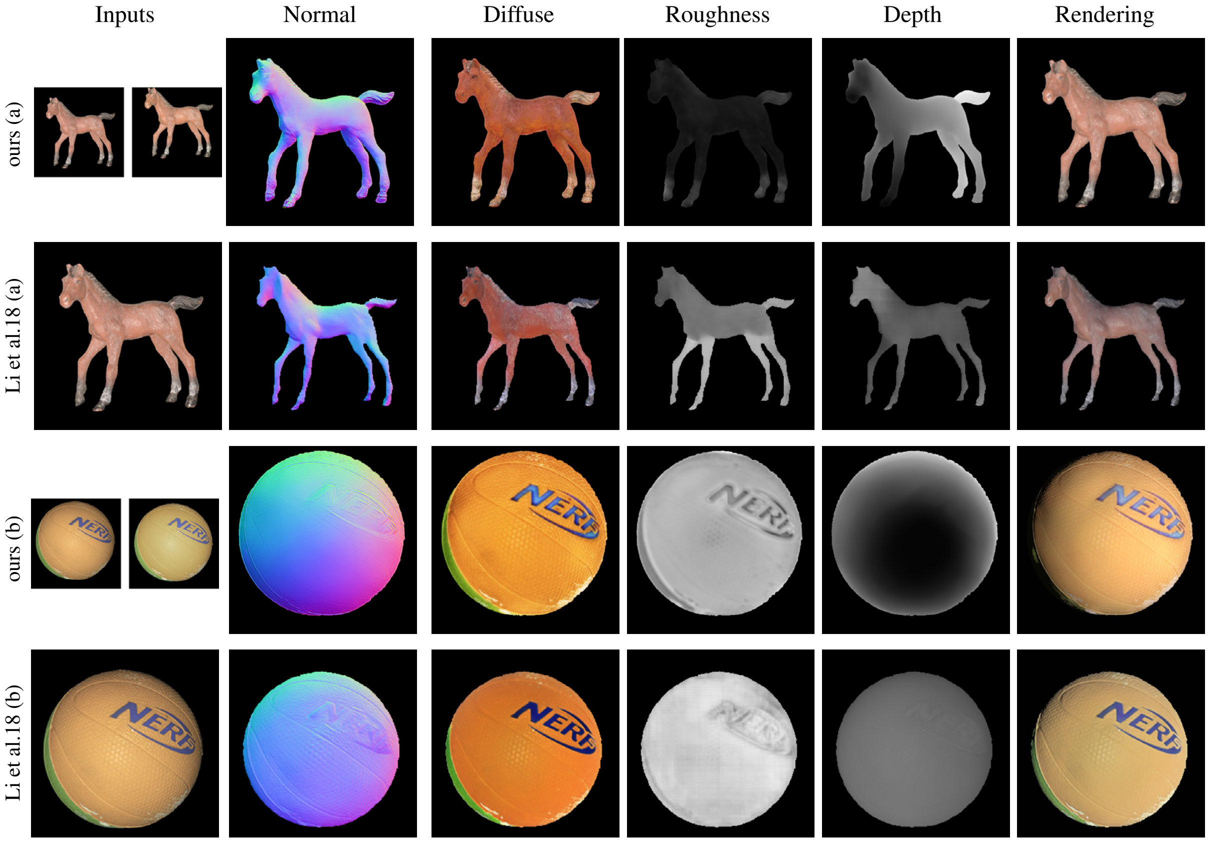


Figure 9: *We compare some real objects to the flash-based method of Li et al. [LXR*18]. Overall, our approach has a much more correct global shape and does not suffer from low-frequency bias and our rendering results are qualitatively better as our shape estimation is superior.*

Numerical figures suggest both our flash-light and natural light results are significantly superior to Li et al. [LXR*18] and Boss et al. [BJK*20]. Our method shows slightly better rendering results on flash data as there is a shading cue from the flash that helps with the SVBRDF estimation, while the method works slightly better on shape estimation with natural lighting. This is because objects are usually better lit under natural lighting conditions, leading to better performance of optical flow. Note that for all quantitative and qualitative comparisons on synthetic data in the paper, we only employ the normals predicted directly by the network and do not perform the detail enhancement step on the normals.

5. Limitation

Our method is limited to a certain-sized object in which the zoom lens sees the whole object and the distance to the object is typically within 50cm. We trained the network on specific data of materials which contains mostly dielectric material samples with few metal examples. Hence the network may not work well on other types of scenes/subjects, e.g. faces or dominantly metallic objects that the network has not seen in the training data (see Figure 11). Metallic objects can also be challenging for optical flow due to bright saturated highlights and texture-less regions. Currently, HDR imaging is required for best results as saturation in images can interfere with optical flow computation, compromising the results. Our network tends to predict slightly blurry surface normals, with high-frequency details in an object's appearance more likely to be attributed as texture in the diffuse albedo. This is due to the normal being constrained by depth estimated from optical flow based correspondence between two views which can be imperfect in textureless regions.

6. Conclusion

In this work, we propose a novel deep learning based method that utilises smartphone multi-lens imaging to estimate shape and SVBRDF for real-world objects. we initially utilize optical flow to estimate the optical flow map between two images. These images are captured with subtle perspective and view shifts, which arise from the multi-lens imaging setup. We then use a trained UNet with a surface rendering loss to estimate the depth from the optical flow map. Finally, the estimated depth is fed together with the target image to the third network to estimate the SVBRDF and surface normal. Our method is relatively robust to incident lighting, making it suitable under both natural and flash illumination. We have demonstrated our object depth estimation method is superior to existing

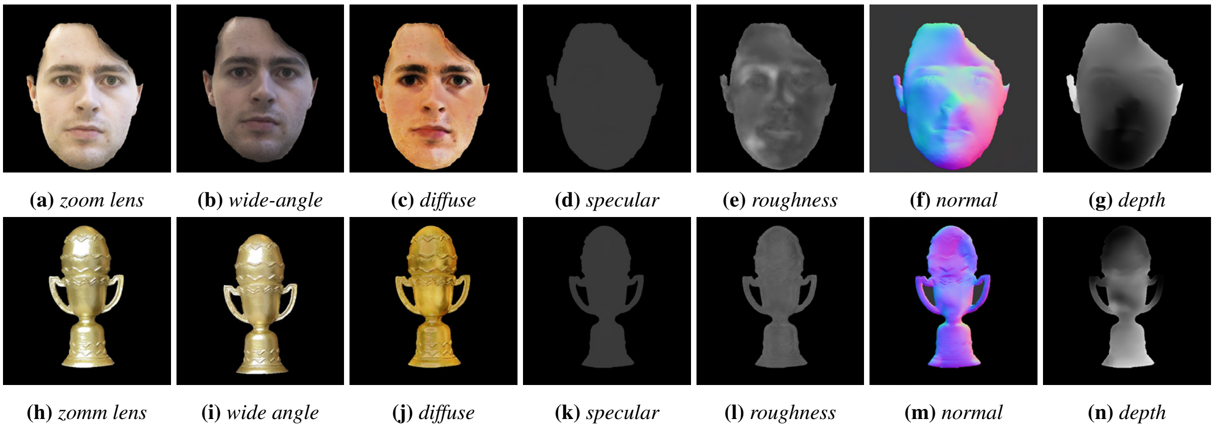


Figure 11: *Some partial failure cases. Since our network has been trained on dominantly dielectric materials, it doesn't work that well on other types of scenes/subjects such as a face (top row), or a dominantly metallic object (bottom row).*

stereo depth methods, and our SVBRDF and surface normal results are overall superior to other state-of-the-art single-image methods that rely on flash illumination. Our method is currently restricted to dominantly dielectric objects as the shape and reflectance estimation is limited by the nature of the synthetic training data.

References

[AAL16] AITTALA, MIIKA, AILA, TIMO, and LEHTINEN, JAAKKO. "Reflectance Modeling by Neural Texture Synthesis". *ACM Trans. Graph.* 35.4 (July 2016), 65:1–65:13. ISSN: 0730-0301 2.

[AWL15] AITTALA, MIIKA, WEYRICH, TIM, and LEHTINEN, JAAKKO. "Two-shot svbrdf capture for stationary materials". *ACM Transactions on Graphics* 34.4 (2015), 110 2.

[BGW*20] BA, YUNHAO, GILBERT, ALEX, WANG, FRANKLIN, et al. "Deep Shape from Polarization". *European Conference on Computer Vision (ECCV)*. 2020 3.

[BJK*20] BOSS, MARK, JAMPANI, VARUN, KIM, KIHWAN, et al. "Two-shot Spatially-varying BRDF and Shape Estimation". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 1, 2, 6–8.

[CBZ*22] CHANG, DI, BOŽIČ, ALJAŽ, ZHANG, TONG, et al. *RC-MVSNet: Unsupervised Multi-View Stereo with Neural Rendering*. 2022. DOI: 10.48550/ARXIV.2203.03949. URL: https://arxiv.org/abs/2203.03949 3.

[DAD*18] DESCHAINTRE, VALENTIN, AITTALA, MIIKA, DURAND, FRÉDO, et al. "Single-Image SVBRDF Capture with a Rendering-Aware Deep Network". *ACM Trans. Graph.* 37.128 (Aug. 2018), 15 2.

[DAD*19] DESCHAINTRE, VALENTIN, AITTALA, MIIKA, DURAND, FRÉDO, et al. "Flexible SVBRDF Capture with a Multi-Image Deep Network". *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)* 38.4 (July 2019). URL: http://www-sop.inria.fr/reves/Basilic/2019/DADDB19 2.

[DDB20] DESCHAINTRE, VALENTIN, DRETTAKIS, GEORGE, and BOUSSEAU, ADRIEN. "Guided Fine-Tuning for Large-Scale Material Transfer". *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)* 39.4 (2020). URL: http://www-sop.inria.fr/reves/Basilic/2020/DDB20 2.

[DLG21] DESCHAINTRE, VALENTIN, LIN, YIMING, and GHOSH, ABHIJEET. "Deep polarization imaging for 3D shape and SVBRDF acquisition". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021 1–4.

[FDI*15] FISCHER, PHILIPP, DOSOVITSKIY, ALEXEY, ILG, EDDY, et al. *FlowNet: Learning Optical Flow with Convolutional Networks*. 2015. DOI: 10.48550/ARXIV.1504.06852. URL: https://arxiv.org/abs/1504.06852 2, 4.

[GLD*19] GAO, DUAN, LI, XIAO, DONG, YUE, et al. "Deep Inverse Rendering for High Resolution SVBRDF Estimation from an Arbitrary Number of Images". *ACM Transactions on Graphics* 37.4 (July 2019). DOI: https://doi.org/10.1145/3306346.3323042 2.

[GRR*17] GEORGOULIS, STAMATIOS, REMATAS, KONSTANTINOS, RITSCHEL, TOBIAS, et al. "Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning". *PAMI* (2017) 2.

[GYY*19] GUO, XIAOYANG, YANG, KAI, YANG, WUKUI, et al. *Group-wise Correlation Stereo Network*. 2019. DOI: 10.48550/ARXIV.1903.04025. URL: https://arxiv.org/abs/1903.04025 2.

[HBNK20] HA, HYUNHO, BAEK, SEUNG-HWAN, NAM, GILJOO, and KIM, MIN H. "Progressive Acquisition of SVBRDF and Shape in Motion". *Computer Graphics Forum* 39.6 (2020), 480–495. DOI: https://doi.org/10.1111/cgf.14087. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14087. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14087 2.

[HJM*22] HWANG, INSEUNG, JEON, DANIEL S., MUOZ, ADOLFO, et al. "Sparse Ellipsometry: Portable Acquisition of Polarimetric SVBRDF and Shape with Unstructured Flash Photography". *ACM Transactions on Graphics (Proc. SIGGRAPH 2022)* 41.4 (2022) 2.

[HSL*17] HUI, ZHUO, SUNKAVALLI, KALYAN, LEE, JOON-YOUNG, et al. "Reflectance Capture Using Univariate Sampling of BRDFs". *ICCV*. Oct. 2017, 5372–5380 2.

[IMS*16] ILG, EDDY, MAYER, NIKOLAUS, SAIKIA, TONMOY, et al. *FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks*. 2016. DOI: 10.48550/ARXIV.1612.01925. URL: https://arxiv.org/abs/1612.01925 2, 4.

[KFR*18] KHAMIS, SAMEH, FANELLO, SEAN, RHEMANN, CHRISTOPH, et al. *StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction*. 2018. DOI: 10.48550/ARXIV.1807.08865. URL: https://arxiv.org/abs/1807.08865 2.

[LDPT17] LI, XIAO, DONG, YUE, PEERS, PIETER, and TONG, XIN. "Modeling Surface Appearance from a Single Photograph Using Self-augmented Convolutional Neural Networks". *ACM Trans. Graph.* 36.4 (July 2017), 45:1–45:11 2.

[LSC18] LI, ZHENGQIN, SUNKAVALLI, KALYAN, and CHANDRAKER, MANMOHAN KRISHNA. "Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image". *ECCV*. 2018 2.

[LTD21] Lipson, Lahav, Teed, Zachary, and Deng, Jia. *RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching*. 2021. doi: 10.48550/ARXIV.2109.07547. url: https://arxiv.org/abs/2109.07547 1–3, 5, 6.

[LWX*22] Li, Jiankun, Wang, Peisen, Xiong, Pengfei, et al. *Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation*. 2022. doi: 10.48550/ARXIV.2203.11483. url: https://arxiv.org/abs/2203.11483 1, 2, 4–6.

[LXR*18] Li, Zhengqin, Xu, Zexiang, Ramamoorthi, Ravi, et al. "Learning to Reconstruct Shape and Spatially-varying Reflectance from a Single Image". *ACM Trans. Graph.* 37.6 (Dec. 2018) 1, 2, 6–8.

[MHG15] Menze, M., Heipke, Christian, and Geiger, Andreas. "JOINT 3D ESTIMATION OF VEHICLES AND SCENE FLOW". *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5 (Aug. 2015), 427–434. doi: 10.5194/isprsannals-II-3-W5-427-2015 2.

[MHP*07] Ma, Wan-Chun, Hawkins, Tim, Peers, Pieter, et al. "Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination". *Proceedings of the 18th Eurographics Conference on Rendering Techniques*. EGSR'07. Grenoble, France: Eurographics Association, 2007, 183–194. isbn: 9783905673524 5.

[MKZ*21] Ma, Xiaohe, Kang, Kaizhang, Zhu, Ruisheng, et al. "Free-Form Scanning of Non-Planar Appearance with Neural Trace Photography". *ACM Trans. Graph.* 40.4 (July 2021) 2.

[MMZ*18] Meka, Abhimitra, Maximov, Maxim, Zollhoefer, Michael, et al. "LIME: Live Intrinsic Material Estimation". *CVPR*. June 2018 2.

[NLGK18] Nam, Giljoo, Lee, Joo Ho, Gutierrez, Diego, and Kim, Min H. "Practical SVBRDF Acquisition of 3D Objects with Unstructured Flash Photography". *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2018)* 36.6 (2018), 267:1–12. doi: 10.1145/3272127.3275017. url: http://dx.doi.org/10.1145/3272127.3275017 2.

[PNS18] Park, Jeong Joon, Newcombe, Richard, and Seitz, Steve. "Surface Light Field Fusion". *2018 International Conference on 3D Vision (3DV)*. 2018, 12–21. doi: 10.1109/3DV.2018.00013 2.

[PSR*17] Pang, Jiahao, Sun, Wenxiu, Ren, Jimmy SJ., et al. *Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching*. 2017. doi: 10.48550/ARXIV.1708.09204. url: https://arxiv.org/abs/1708.09204 2.

[RPG16] Riviere, J., Peers, P., and Ghosh, A. "Mobile Surface Reflectometry". *Computer Graphics Forum* 35.1 (2016), 191–202. issn: 1467-8659 2, 5.

[SHK*14] Scharstein, Daniel, Hirschmüller, Heiko, Kitajima, York, et al. "High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth". Vol. 8753. Sept. 2014, 31–42. isbn: 978-3-319-11751-5. doi: 10.1007/978-3-319-11752-2_3 2.

[SSG*17] Schoeps, Thomas, Schonberger, Johannes, Galliani, Silvano, et al. "A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos". July 2017. doi: 10.1109/CVPR.2017.272 2.

[TD20] Teed, Zachary and Deng, Jia. *RAFT: Recurrent All-Pairs Field Transforms for Optical Flow*. 2020. doi: 10.48550/ARXIV.2003.12039. url: https://arxiv.org/abs/2003.12039 1–5.

[THZ*20] Tankovich, Vladimir, Häne, Christian, Zhang, Yinda, et al. *HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching*. 2020. doi: 10.48550/ARXIV.2007.12140. url: https://arxiv.org/abs/2007.12140 2.

[WWZ16] Wu, Hongzhi, Wang, Zhaotian, and Zhou, Kun. "Simultaneous Localization and Appearance Estimation with a Consumer RGB-D Camera". *IEEE Transactions on Visualization and Computer Graphics* 22.8 (2016), 2012–2023. doi: 10.1109/TVCG.2015.2498617 2.

[WZ15] Wu, Hongzhi and Zhou, Kun. "AppFusion: Interactive Appearance Acquisition Using a Kinect Sensor". *Computer Graphics Forum* 34.6 (2015), 289–298. issn: 1467-8659 2.

[XLZ*23] Xu, Xianmin, Lin, Yuxin, Zhou, Haoyang, et al. "A Unified Spatial-Angular Structured Light for Single-View Acquisition of Shape and Reflectance". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 206–215 2.

[XZ20] Xu, Haofei and Zhang, Juyong. *AANet: Adaptive Aggregation Network for Efficient Stereo Matching*. 2020. doi: 10.48550/ARXIV.2004.09548. url: https://arxiv.org/abs/2004.09548 1, 2, 4, 6.

[XZC*22] Xu, Haofei, Zhang, Jing, Cai, Jianfei, et al. *Unifying Flow, Stereo and Depth Estimation*. 2022. doi: 10.48550/ARXIV.2211.05783. url: https://arxiv.org/abs/2211.05783 2, 5.

[YLD*18] Ye, Wenjie, Li, Xiao, Dong, Yue, et al. "Single Photograph Surface Appearance Modeling with Self-Augmented CNNs and Inexact Supervision". *Computer Graphics Forum* 37.7 (Oct. 2018). doi: https://doi.org/10.1111/cgf.13560 2.

[YMHR19] Yang, Gengshan, Manela, Joshua, Happold, Michael, and Ramanan, Deva. *Hierarchical Deep Stereo Matching on High-resolution Images*. 2019. doi: 10.48550/ARXIV.1912.06704. url: https://arxiv.org/abs/1912.06704 2.

[ZYR*22] Zhang, Jinrui, Yang, Huan, Ren, Ju, et al. "MobiDepth: Real-Time Depth Estimation Using on-Device Dual Cameras". *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. MobiCom '22. Sydney, NSW, Australia: Association for Computing Machinery, 2022, 528–541. isbn: 9781450391818. doi: 10.1145/3495243.3560517. url: https://doi.org/10.1145/3495243.3560517 2.

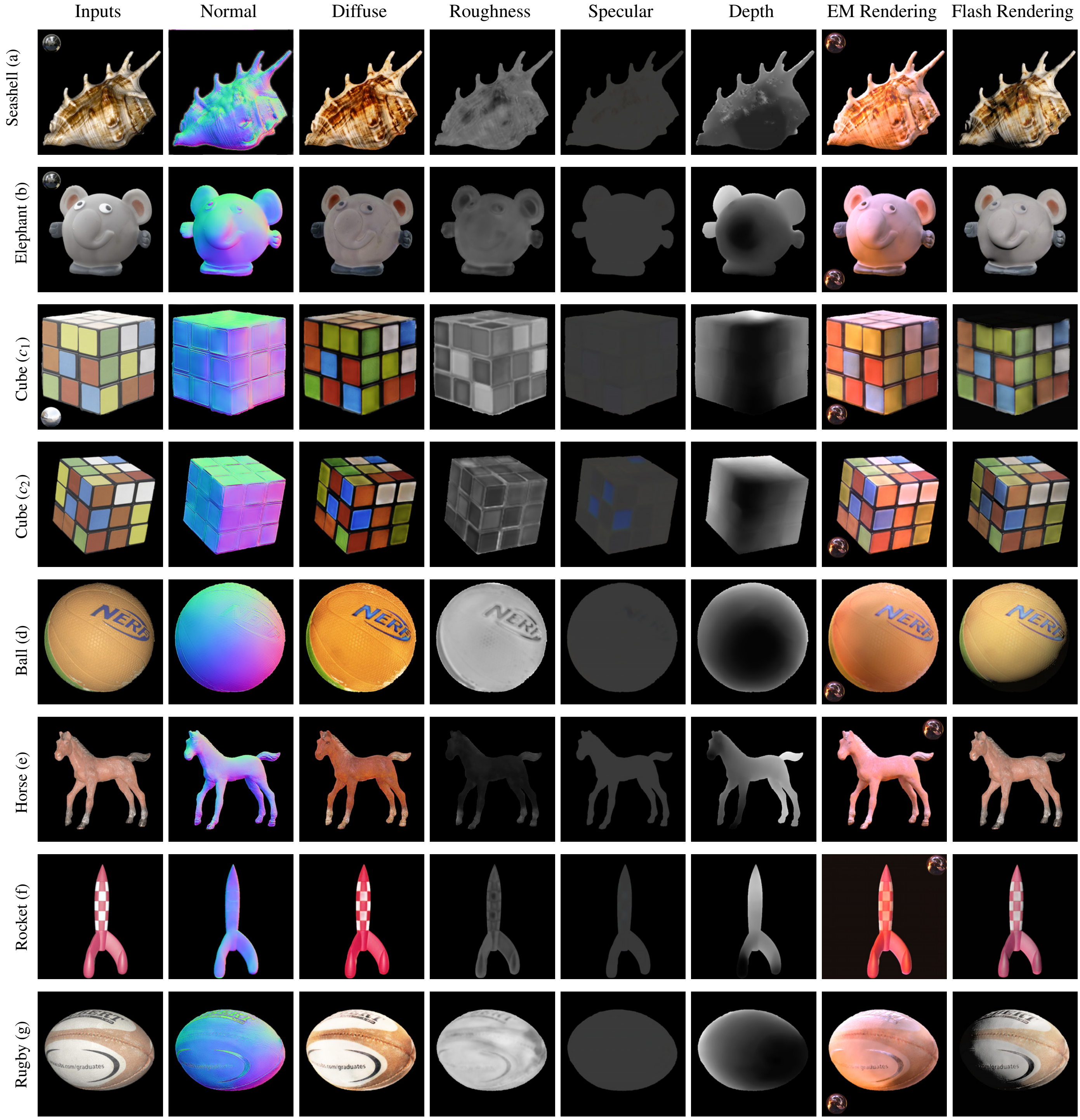


Figure 12: *We tested our method on various real-world 3D objects under both natural lighting conditions (a, b, c_1) and flashlights (c_2, d, e, f, g). Our estimated shape and reflectance maps generate high-quality rendering results under both environmental illumination (Grace cathedral), and flash illumination.*

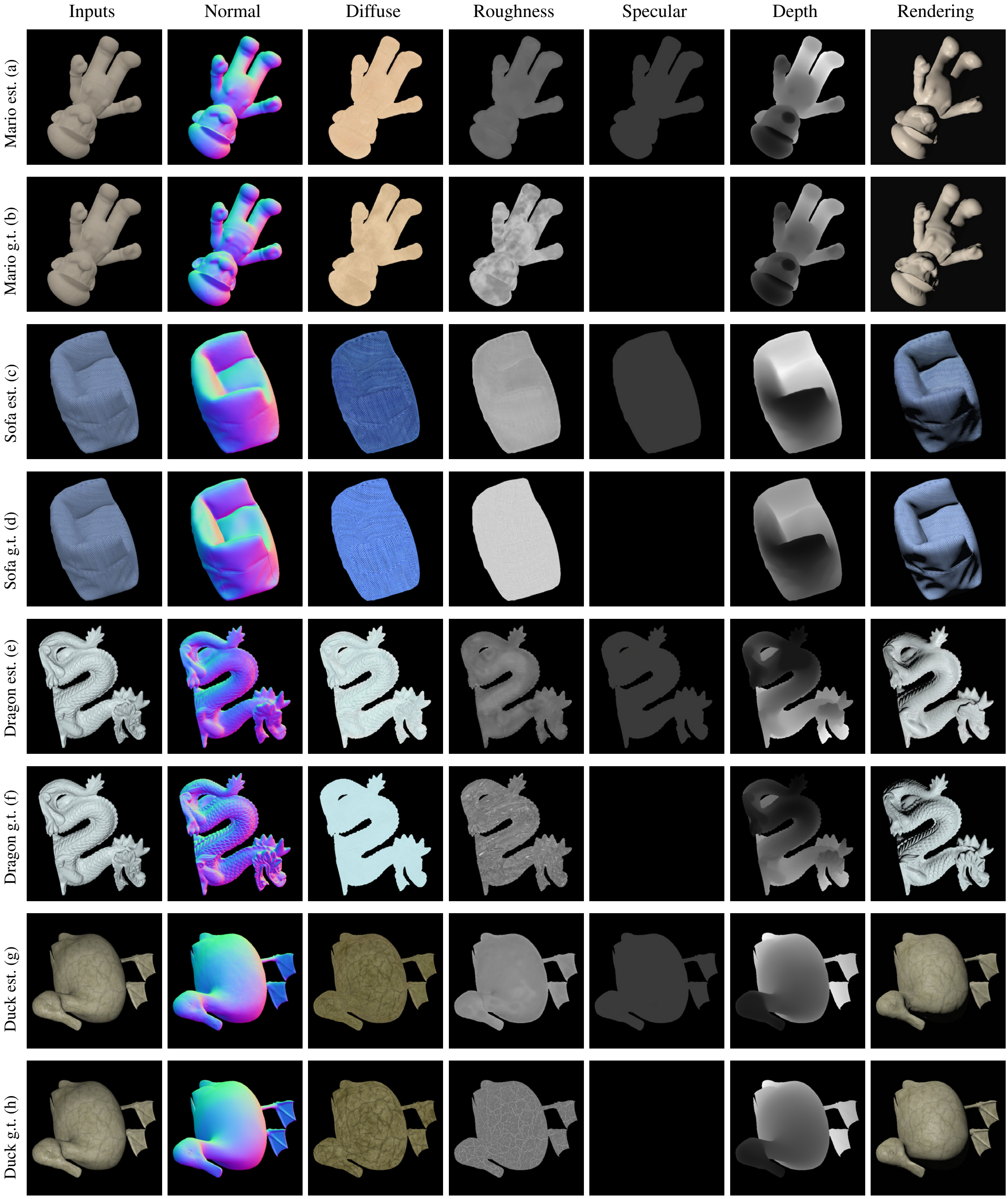


Figure 13: *We compare our estimated results (a, c, e, g) to g.t. (b, d, f, h) on various synthetic 3D objects not seen during training, under both natural lighting conditions (a, c) and flashlights (e, g). Our estimated maps are accurate in comparison to g.t. maps, and renderings are close to g.t. renderings.*