Multi-View Flash and No-Flash Photogrammetry for Shape Recovery

TINA TIAN, Carnegie Mellon University, USA

This paper presents a Multi-View Photometric Stereo (MVPS) pipeline designed for surface normal estimation, utilizing a hardware setup comprising only a monocular camera and a flashlight. The study involves the reimplementation of a proprietary Lambertian object shape refinement technique, employing both flash and no-flash imaging methodologies. Furthermore, we integrate this shape refinement routine with an open-source Structure-from-Motion (SfM) method. Evaluation results demonstrate a qualitative enhancement of the normal map by introducing higher-frequency components, contrasting with the coarse normals generated by SfM alone, albeit with a slightly increased mean angular error in the refined normals. We conduct an analysis to explore potential reasons for this performance discrepancy. Despite the encountered challenges, the project yields valuable outcomes, including the implementation of fundamental MVPS procedures and the creation of a tailored simulation pipeline for data capture and ground-truth generation. The envisaged utility of this simulation pipeline extends to future projects in the field.

ACM Reference Format:
Tina Tian. 2023. Multi-View Flash and No-Flash Photogrammetry for Shape Recovery. 1, 1 (December 2023), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Shape recovery from images is a fundamental problem in computer vision, with applications in scene reconstruction for robot manipulation, navigation, inspection, augmented reality, and many others. There are various schools of thought for shape recovery using images. Two of the most well-adopted methods are geometric and photometric approaches (Fig. 1), and each has different pros and cons. Geometric approaches take images of a scene from multiple viewpoints, find point correspondences across images, and establish their geometric position to recover the shape. Examples of this type of approach include structure from motion (SfM) and stereo imaging. Though requiring relatively simple capturing devices, conventional geometric-based methods have fundamental limitations in recovering high-frequency details [Klowsky et al. 2012]. Photometric methods, on the other hand, recover per-pixel surface orientation using shading cues instead of disparity between views. This type of approach typically relies on multiple images captured under different lighting conditions which can be hard to set up [Yuille and Snow 1997].

Due to the complementary characteristics of geometric and photometric-based approaches, it is interesting and natural to envision a method that amalgamates geometric and photographic approaches to obtain high-quality shape recovery. It turns out that this task, termed multi-view photometric stereo (MVPS), is not new. It is often formulated as SfM plus an energy optimization problem, solved iteratively from a coarse initialization [Cao et al. 2020]. In this research, we present an MVPS pipeline that performs surface normal estimation leveraging a minimal handheld hardware setup, consisting only of a monocular camera and a flashlight. We focus on implementing the closed-source method for Lambertian object shape refinement based on flash and no-flashing imaging [Cao et al. 2020] and combining it with an open-source state-of-the-art SfM method [Sarlin et al. 2019].

To evaluate our method, we develop a customized simulation pipeline for data capturing and ground-truth generation. Results show that our approach recovers a normal map with more high-frequency components compared to the coarse normals from SfM alone. Interestingly, however, the mean angular error of refined normals is higher. We also experiment using real-world data to demonstrate the functionality of our algorithm.

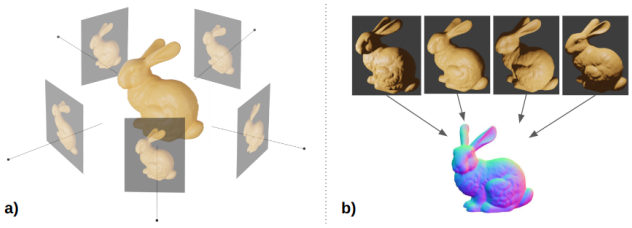

Fig. 1. (a) Geometric and (b) photometric approaches for image-based 3D reconstruction.

2 RELATED WORK

Multi-view Stereo (MVS) has long been used to reconstruct 3D shapes by combining information from multiple images. Fundamental to multi-view stereo is the understanding of epipolar geometry and accurate camera calibration. Ensuring precise calibration parameters and establishing correspondences between image points are crucial for accurate 3D reconstruction. The process of establishing correspondences between images involves feature matching techniques. Keypoint extraction and matching algorithms, such as SIFT (Scale-Invariant Feature Transform) and SURF (Speeded Up Robust Features), are widely employed in this context. One of the most widely-used MVS methods, COLMAP [Schönberger and Frahm 2016; Schönberger et al. 2016], performs the feature extraction and matching using the conventional SIFT feature descriptor, and it struggles to find correct matches under feature-sparse scenarios or varying lighting conditions. To address this issue, learning-based feature extraction [DeTone et al. 2017] and matching [Sarlin et al. 2020] has been proposed and have achieved superior results. Sarlin et al. propose a hierarchical localization approach that utilizes

Author's address: Tina Tian, yutian@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, USA, 15213.

ACM XXXX-XXXX/2023/12-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

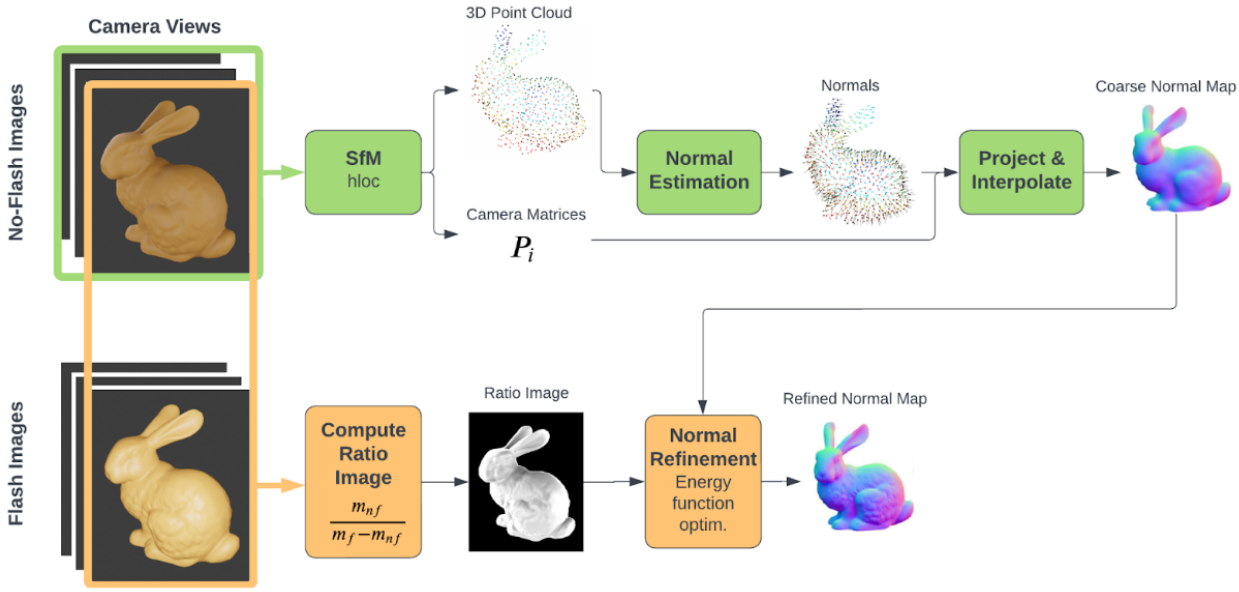

Fig. 2. The proposed architecture. The no-flash images are first used to generate a coarse 3D point cloud reconstruction using structure from motion (SfM) and recover a coarse normal map. Then, each coarse normal map is refined through energy optimization using shading cues from the corresponding flash and no-flash image pair.

learned descriptors, achieving remarkable localization robustness [Sarlin et al. 2019]. In this paper, we make use of their open-source toolbox.

Flash and no-flash imaging represent two fundamental approaches to scene illumination. Flash photography involves the use of artificial light sources to illuminate scenes, while no-flash imaging relies solely on ambient light. Conventionally, flash and no-flash photography are performed separately and have complementary properties. Flash photography is often used in low-light conditions such as nighttime, or in portraiture to highlight facial features and reduce shadows. Though capable of capturing high-frequency details in the scene, direct flash can create harsh shadows, affecting the aesthetics of the image. No-flash photography is also used to capture images in low-light conditions without artificial illumination, and the purpose is, contrary to flash photography, to capture the ambiance of the scene. Due to the limited available light, the no-flash image can have increased image noise.

[Petschnigg et al. 2004] is one of the earliest work that integrates flash and no-flash techniques. In this work, the authors show that under low ambient light, combining flash/no-flash image pairs can preserve the ambient qualities of the no-flash image as well as the high-frequency flash detail in the flash image, and this has applications in image denoising and detail transfer. When comparing a flash and a no-flash image, another observation is that objects close to the flashlight have a stronger change in appearance than distant objects. This property has been used in image matting [Sun et al. 2006] and foreground extraction [Sun et al. 2007].

Leveraging the complementary characteristics of flash and no-flash imaging, in [Zhou et al. 2012] the authors propose a method that uses flash and no-flash imaging to refine shape reconstruction. They show that the ratio of a flash/no-flash pair can make stereo matching robust against depth discontinuities. Building on top of this idea, [Cao et al. 2020] use this ratio image to formulate shading constraints based on the image formation model for Lambertian objects. This constraint is then used in an energy optimization step, which refines the coarse depth and normal generated from stereo matching. They have shown that the refined normal and depth have lower errors compared to the coarse counterparts. Other work addresses the shape recovery problem for objects with more challenging material. For example, [Cheng et al. 2021] proposes a method to recover the shape of reflective objects using a point light source attached to a moving camera. A lighting setup similar to [Cheng et al. 2021] is used in this project, though our focus are Lambertian objects.

3 METHODOLOGY

Fig. 2 depicts the overall process of our coarse-to-fine shape recovery approach. We first capture a set of flash/no-flash image pairs at different camera views. Then, we derive the rough shape through multi-view stereo using no-flash images captured at multiple views [Sarlin et al. 2019] [Sarlin et al. 2020]. Subsequently, we employ the flash/no-flash pair to refine the coarse shape [Cao et al. 2020], acquiring a higher-resolution shape. This coarse-to-fine shape recovery is facilitated by our image formation model, which is established based on the flash/no-flash pair.

3.1 Geometry-Based Coarse Shape Recovery

To recover the coarse shape, we first extract features using learned descriptor and feature matching methods [DeTone et al. 2017; Sarlin et al. 2020], then perform hierarchical multi-view stereo [Sarlin et al. 2019]. The output is a noisy point cloud, which is then post-processed to denoise by applying the Statistical Outlier Removal (SOR) filter multiple times. This process is manually performed because the noise distribution and amount varies from scene to scene. Then, we estimate the normal for each point and perform Poisson surface reconstruction [Kazhdan et al. 2006]. The normal map is obtained by projecting the reconstructed mesh to the image sensor using the estimated camera matrix.

3.2 Image Formation

We now detail the process of photometric shape refinement using hybrid flash and no-flash imaging. We employ the same image formation as [Cao et al. 2020]. The pixel intensities of the flash and no-flash images, m_f and m_{nf}, are formulated as the product of an albedo $\rho \in \mathbb{R}$ and a shading function $s : \mathcal{S}^2 \to \mathbb{R}$, as shown in (1):

$$\begin{cases} m_{\text{nf}} = \rho s_{\text{nf}}(\mathbf{n}) \\ m_{\text{f}} - m_{\text{nf}} = \rho s_{\text{fo}}(\mathbf{n}), \end{cases} \tag{1}$$

The no-flash shading function $s_{nf}(\mathbf{n})$ and the flash-only shading function $s_{fo}(\mathbf{n})$ are different. For Lambertian objects, at a surface patch with normal $\mathbf{l} \in \mathcal{S}^2 \subset \mathbb{R}^3$ with intensity $e : \mathcal{S}^2 \to \mathbb{R}$ given by (2). This shading function holds for point-light scenarios.

$$s(\mathbf{n}) = e(\mathbf{l})\, \mathbf{n}^\top \mathbf{l}, \tag{2}$$

We assume that our flashlight is a point light at infinity, it is aligned with the camera's principal axis and pointing towards the camera (the flashlight direction is $[0, 0, -1]^\top$). Applying the shading function (2), we get (3), where e_f is the flash intensity:

$$s_{fo}(\mathbf{n}) = e_f\, [n_1, n_2, n_3] [0, 0, -1]^\top = -e_f\, \mathbf{n}_3 \tag{3}$$

In the no-flash image, on the other hand, light rays reach the surface patch from multiple directions, and the reflected light becomes the integral over all possible incident directions, as shown in (4):

$$s_{nf}(\mathbf{n}) = \int_{\mathcal{S}^2} e(\mathbf{l})\, \mathbf{n}^\top \mathbf{l}\, d\mathbf{l} \tag{4}$$

We linearize this equation following [Basri and Jacobs 2003], which proves that for a Lambertian surface, its reflectance can be modeled as a low-dimensional linear combination of spherical harmonics. We use second-order spherical harmonics to parameterize the shading function under no-flash conditions. Denote $n = [n_1, n_2, n_3]^\top$, second-order spherical harmonics can be stacked into a vector $h(\mathbf{n})$ as shown in (5). Then, the linearization of (4) becomes (6), where $\mathbf{k}_{nf} \in \mathbb{R}^9$ is a stack of linear combination coefficients.

$$h(n) = [1, n_1, n_2, n_3, n_1 n_2, n_1 n_3, n_2 n_3, n_1^2 - n_2^2, 3n_2^2 - 1]^\top \tag{5}$$

$$s_{nf}(\mathbf{n}) = h(\mathbf{n})^\top \mathbf{k}_{nf} \tag{6}$$

Dividing the two equations in (1) and substituting the shading functions with (3) and (6), we get the *ratio image*, where the albedo ρ is canceled out, as shown in (7):

$$\frac{h(\mathbf{n})^\top \mathbf{k}'}{-n_3} = \frac{m_{nf}}{m_f - m_{nf}}, \tag{7}$$

Here $\mathbf{k}' = \mathbf{k}_{nf}/e_f$ is the *global lighting vector*, which is the spherical harmonics coefficient vector scaled by flashlight intensity. This image formation model now explicitly relates surface normal, lighting, and observed intensity.

3.3 Shape Refinement Through Optimization

We now detail the shape refinement process given a flash/no-flash pair and the coarse shape. As in [Cao et al. 2020], we formulate the surface normal recovery as per-pixel energy function optimization with a shading constraint, a surface normal constraint, and a unit-length constraint:

$$\min_{\mathbf{n}} \omega E_s(\mathbf{n}) + \lambda_1 E_n(\mathbf{n}) + \lambda_2 E_u(\mathbf{n}), \tag{8}$$

where ω denotes the confidence of the image formation model at a pixel, $r = m_f/m_{nf}$, and μ and σ are the mean and standard deviation of all pixels in r, as shown in (9):

$$\omega = \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right) \tag{9}$$

The weight ω downweights cast shadows–shadows projected by the object in the direction of the light source–which are not modeled in the image formation described in (3). What this simple shading model does handle are the attached shadows, which exist in the part of the object that is not illuminated by direct light [Salvador et al. 2001]. The use of ω is motivated by the observation that cast shadows strongly increase the ratio r while highlights strongly decrease the ratio, making the ratio deviate far from the mean ratio. An example of this phenomenon is shown in Fig. 3, where the rim of the leg has high values in the ratio image due to cast shadow.

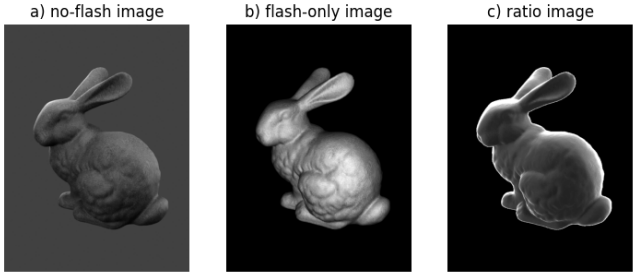

Fig. 3. Comparison of (a) no-flash, (b) flash-only, and (c) ratio image.

For now, assume the global lighting vector $\mathbf{k}'$ is known. Then (7) is with only three unknowns, which are the elements in $\mathbf{n}$. Thus, we write (7) as a shading constraint on $\mathbf{n}$, as shown in (10):

$$E_s(\mathbf{n}) = \left(h(\mathbf{n})^\top \mathbf{k}' + n_3 \frac{m_{nf}}{m_f - m_{nf}}\right)^2 \tag{10}$$

Two other constraints, the normal constraint (11) and the unit length constraint (12), state that the estimated normal should be

close to the coarse normal and that the estimated normal should have unit length, respectively:

$$E_n(\mathbf{n}) = \left(1 - \mathbf{n}^\top \mathbf{n}^{(0)}\right)^2 \tag{11}$$

$$E_u(\mathbf{n}) = \left(1 - \mathbf{n}^\top \mathbf{n}\right)^2. \tag{12}$$

Before actually performing the energy optimization in (8), we need to compute the missing piece $\mathbf{k}'$. To do so, we bootstrap from the initial guess $\mathbf{n}^{(0)}$, which is the coarse normal map we obtained from SfM. Suppose there are p pixels in the region of interest, for each pixel we stack the row vector $h(n)^\top/(-n_3)$ vertically into a matrix $N \in \mathbb{R}^{p\times 9}$ and stack all observed $m_{nf}/(m_f - m_{nf})$ into a vector $\mathbf{m} \in \mathbb{R}^p$, yielding an overdetermined linear system:

$$N\,\mathbf{k}' = \mathbf{m}. \tag{13}$$

In practice, we find that performing the aforementioned optimization on each pixel of an image (e.g. size 800×600) is very time-consuming. To speed up the computation, in our implementation, the optimization is parallelized in the following way:

- The image is patchified into neighborhoods with size 25×25.
- For each patch, optimize the normal of each pixel sequentially.
- The optimization job of all patches is computed using multiprocessing.

4 EXPERIMENTS AND RESULTS

In this section, we show experiments on both synthetic and real-world data to evaluate our method.

4.1 Synthetic Data

We evaluate our method both qualitatively and quantitatively using synthetic data. The data is generated using Blender with physics-based renderer Cycles. Two publically available 3D objects, Stanford Bunny and Happy Buddha [rep [n. d.]], are rendered with Lambertian materials under perspective projection. In our simulation, the background is set to a uniform color. We simulate no-flash images using four area lights. To simulate flash images, we put a point light on the camera center. Instead of simulating the camera and the light assembly moving over time, we create multiple such assemblies, each with a different view. We use a total of 122 views. Fig. 4 shows some examples of no-flash images. During coarse shape recovery, it takes roughly 15 minutes to complete the feature extraction, matching, and reconstruction steps. In shape refinement, it takes 35 seconds per image to complete the optimization.

To quantitatively evaluate our method, we export the ground truth normal map and depth map of each camera view and compare them against the estimated normal and depth maps.

Shape Recovery Results Fig. 5 shows examples of the shape recovery results along with their coarse initializations and ground truth. For Eq.(8) in normal refinement, the parameters λ_1, λ_2 are both set to 0.1. The metric we use is the mean angular error (MAngE) of the normal vectors in the foreground across 4 different views. Qualitatively speaking, the coarse normal map contains only a low-poly geometry, while the normal map refined using our implementation contains slightly more details. For example, in the bunny's eye, ear, and back leg regions as well as the buddha's face and sleeve regions, the refined normal maps have more pronounced bumps and grooves, which look more similar to the true normal maps. However, our quantitative evaluation shows that the refined normal map returned from our energy optimization has a higher MAngE than the coarse normal map we used to bootstrap the optimization (Table 1, 2). This result is unexpected since we are using the same parameters as mentioned in the reference paper [Cao et al. 2020].

The exact cause of the increase in MAngE is yet to be verified. Here are a few guesses on why the refined normals have higher errors on average. First, the lighting conditions used in the reference paper are not completely the same as our simulation. It might be the case that the strength of the flashlight affects the effectiveness of the flash and no-flash imaging-based normal refinement because

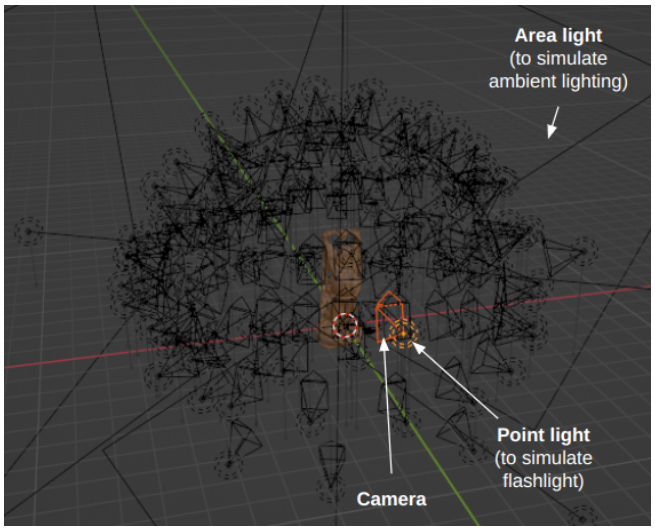

Fig. 4. Examples of Happy Buddha no-flash images used in geometry-based coarse shape recovery.

Table 1. Mean Angular Error of Coarse and Refined Normal Maps of Stanford Bunny

Type	Mean Err	Max Err	Min Err
Coarse	**0.17954**	**2.35271**	**0.00054**
Refined	0.23038	2.71571	0.00061

Table 2. Mean Angular Error of Coarse and Refined Normal Maps of Happy Buddha

Type	Mean Err	Max Err	Min Err
Coarse	**0.49948**	**2.51896**	**0.00059**
Refined	0.50160	2.83051	0.00091
Refined (w/o normal constraint)	0.53590	3.04205	0.00107
Refined (with true k′)	0.50850	2.85146	0.00108

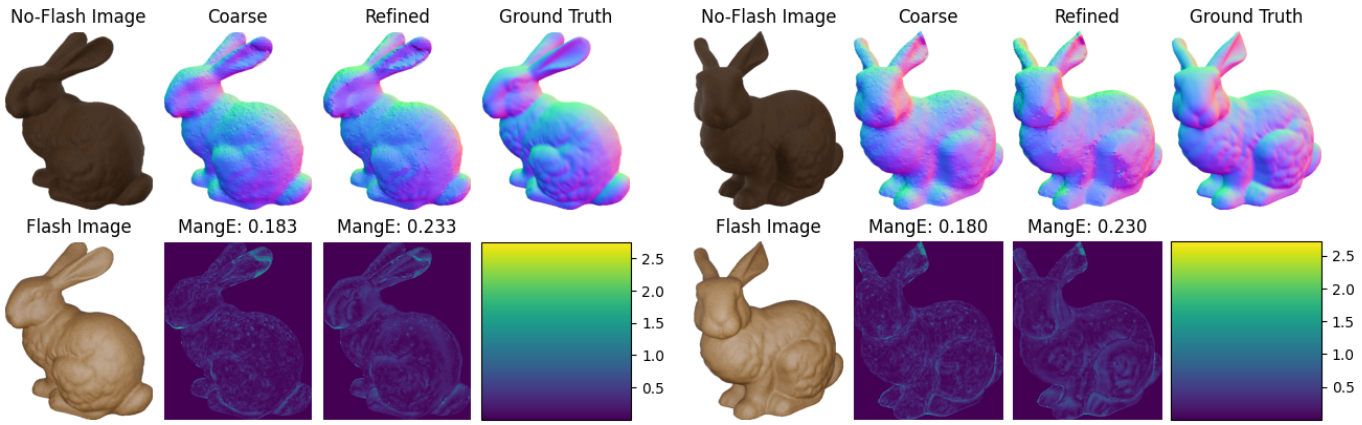

Fig. 5. Shape recovery results on synthetic data (Stanford Bunny). The first column shows the rendered flash/no-flash pair. The first row shows the coarse, refined, and true normal maps. The second row displays the error maps. The number above each error map is the MAngE of the corresponding normal map. Although our refined normal map appears to have more high-frequency details compared to the coarse normal maps, the MAngE of the refined normal map is higher than that of the coarse normal map.

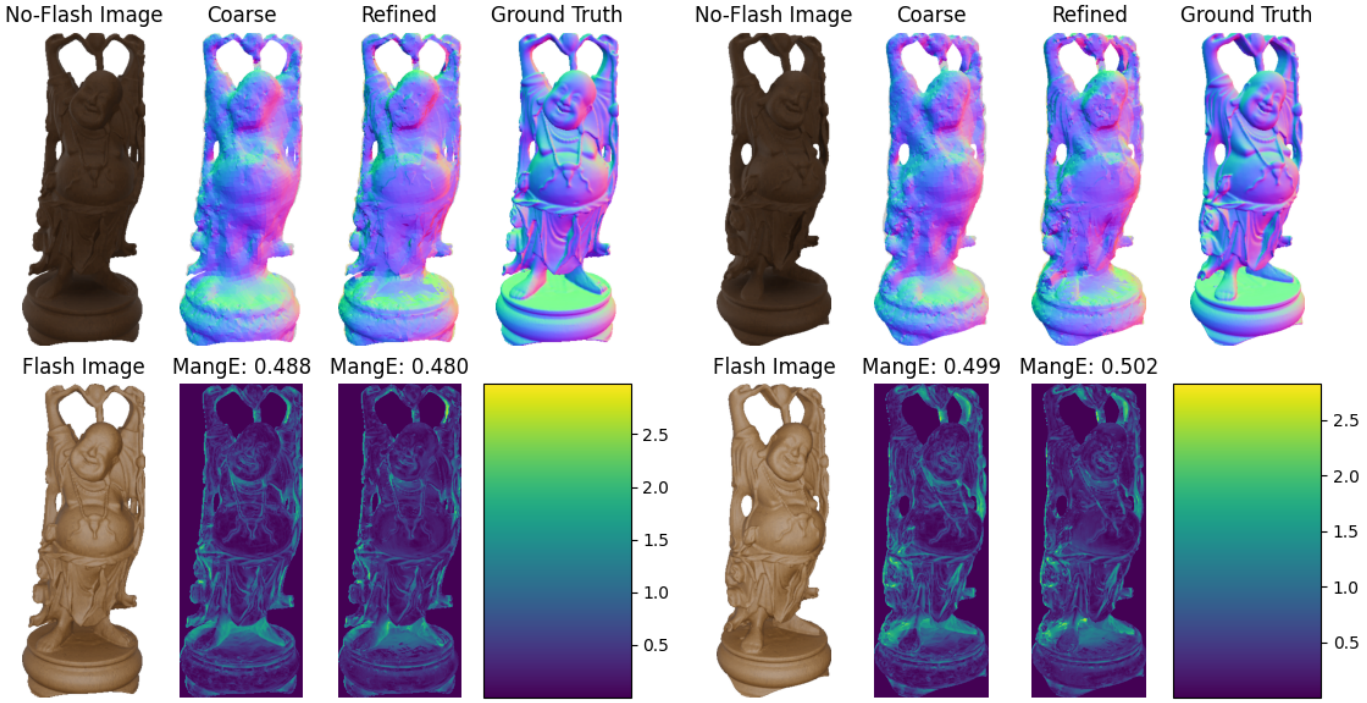

Fig. 6. Shape recovery results on synthetic data (Happy Buddha).

a stronger flashlight could darken the cast shadow and brighten the highlighted area, leading to a change in the shading confidence weight ω in (9) and thus a change in the energy function. Another possible cause is the different materials used in our implementation and the reference paper. More experiments need to be conducted to investigate the effect of lighting and material on the algorithm performance.

Below we analyze two other possibilities that might be related to the unsatisfactory performance.

Effect of Normal Constraint By observing the difference between the coarse and the refined normal maps, it is tempting to guess that the incorrect coarse normal could propagate to the refined normal due to the normal constraint. To see whether this is the case, after we bootstrap the global lighting vector $\mathbf{k}'$ in (13)

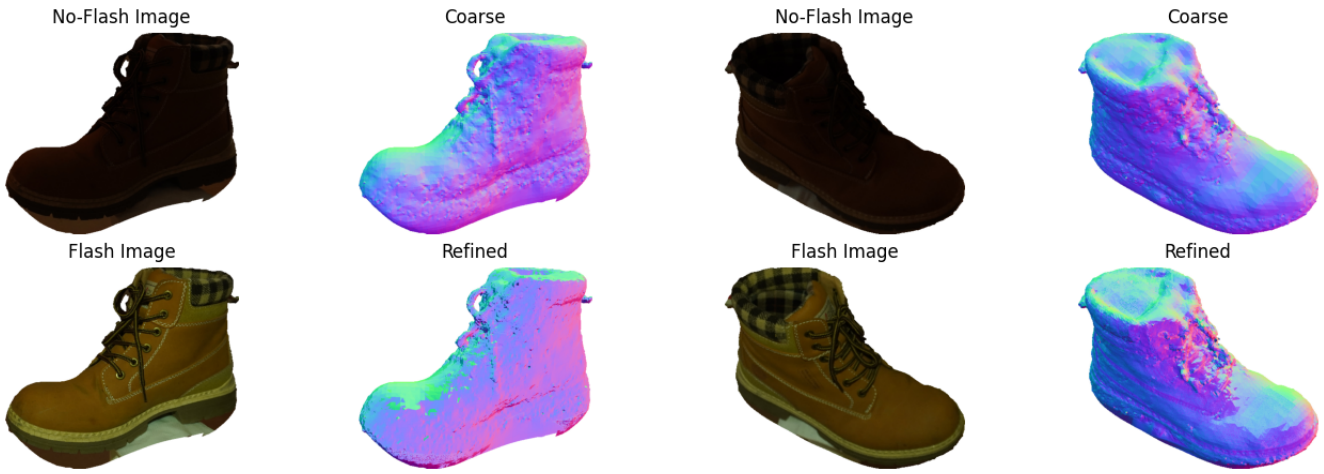

Fig. 7. Normal maps of a real-world object (a boot). The refined normal map contains artifacts at the surface, which is likely caused by the incorrect identification of shadow.

and proceed with the optimization in (8), we disable the normal constraint by setting $\lambda_1 = 0$. The results are shown in Table 2. The normal estimation accuracy is lower compared to the case when the normal constraint is enabled. This means that the normal constraint does help the normal refinement.

Effect of Global Lighting Vector Our normal refinement method relies on the computation of the global lighting vector $\mathbf{k}'$ to bootstrap the optimization. Experiments show that without a roughly correct $\mathbf{k}'$ (e.g. by initializing the coarse normal map to random), the result generated from the optimization becomes meaningless. To get a roughly correct estimation of $\mathbf{k}'$, according to (13), it is required to provide a coarse normal map $\mathbf{n}^{(0)}$. For this reason, if the normal initialization is too different from the true normal, $\mathbf{k}'$ will also be off and the optimization will be unfounded.

To see if the coarse normal produces an incorrect global lighting vector that results in a wrong optimization, we take the ground truth normal map to compute the true global lighting vector as in (13), and then use it in the shading constraint (1). As shown in Table 2, using the true global lighting vector, there is no improvement to the refined normal error. This means that the impact of global lighting vector error on the refinement performance is minimal, given that our coarse normal is roughly correct.

4.2 Real-World Experiments

Experiment Setup Fig. 8 shows our custom image capturing setup. This device has a monocular camera with 2MP 1/2.9" OG02B10 color global shutter sensor, paired with 38.92mm lens. Three LED beads form the flashlight and they are placed around the camera, facing the viewing direction. An onboard computer (Raspberry Pi Zero 2W) synchronizes the shutter and the lights. The light is quickly and constantly strobing, such that the flash and no-flash frames are captured alternatingly, with a close to zero time gap. This design is meant to be handheld and eliminates the need to use a tripod. With custom data streaming software, it can communicate with a base station computer wirelessly and visualize the images in real time.

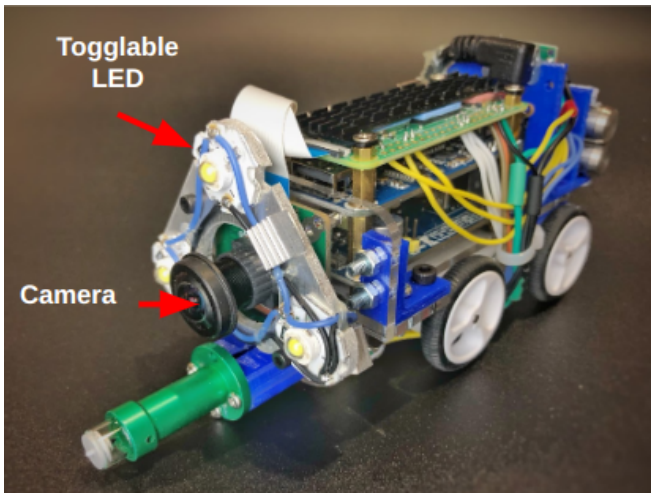

Fig. 8. Physical data capturing device.

We capture 40 flash/no-flash image pairs of a boot at different angles. Examples of the captured images are shown in Fig. **??**. The coarse normal map and the refined normal map are shown in 7. As we can see from the results, the refined normal maps have more artifacts than the coarse normal map. For example, the instep of the boot is less smooth after the normal refinement, and the grooves in the coarse normal map are smeared in the refined normal maps. This means that the normal refinement not only does not recover more high-frequency components in this case but also aggravates the error of the normal estimation. Considering the lack of smoothness of the refined normal map, one thing that could possibly improve the performance is to include an additional smoothness term in the energy function (8), which will make the cost of each pixel not only

a function of the pixel itself but also the context around it. This will likely complicate the optimization process and make the parallel optimization more challenging.

5 CONCLUSION

In this work, we present an MVPS pipeline designed for surface normal estimation, utilizing a minimal hardware configuration comprising solely a monocular camera and a flashlight. One part of our work is a full implementation of the closed-source technique for Lambertian object shape refinement based on flash and no-flash imaging [Cao et al. 2020]. Another part of our work is the integration of the shape refinement routine with an open-source Structure-from-Motion (SfM) method [Sarlin et al. 2019]. The evaluation results demonstrate that our approach qualitatively enhances the normal map by incorporating more high-frequency components compared to the coarse normals generated by SfM alone. However, it is noteworthy that the mean angular error of the refined normals is observed to be slightly higher. Due to the limited time frame of this project, the cause of the undesirable performance of our approach is still unknown.

Regardless, there are important takeaways from this project. While we had limited prior experience with Multi-view Photometric Stereo (MVPS) and dataset collection in a simulator like Blender, during this project, we had a chance to implement the core procedures of Multi-view Photometric Stereo (MVPS) and develop a custom simulation pipeline for data capture and ground-truth generation. We believe this simulation pipeline will be helpful for future projects as well.

Currently, our normal refinement pipeline is capable of recovering the normal map of each camera view using a pair of flash/no-flash images, and each normal map is only a partial reconstruction of the full object. It will be interesting to compute the refined normals of all views and merge them into a full 3D reconstruction. This could be another research and requires conducting a literature review on normal map merging. Also, to compute all normal maps for different views, algorithmic optimization needs to be performed to speed up the computation, possibly by enabling GPU computing, so that the reconstruction can be completed within a reasonable amount of time.

REFERENCES

[n. d.]. https://graphics.stanford.edu/data/3Dscanrep/

R. Basri and D.W. Jacobs. 2003. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 2 (2003), 218–233.

Xu Cao, Michael Waechter, Boxin Shi, Ye Gao, Bo Zheng, and Yasuyuki Matsushita. 2020. Stereoscopic Flash and No-Flash Photography for Shape and Albedo Recovery. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3427–3436.

Ziang Cheng, Hongdong Li, Yuta Asano, Yinqiang Zheng, and Imari Sato. 2021. Multi-view 3D Reconstruction of a Texture-less Smooth Surface of Unknown Generic Reflectance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16226–16235.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2017. SuperPoint: Self-Supervised Interest Point Detection and Description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), 337–33712.

Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. 2006. Poisson Surface Reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*. 61–70.

Ronny Klowsky, Arjan Kuijper, and Michael Goesele. 2012. Modulation transfer function of patch-based stereo systems. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1386–1393.

G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. 2004. Digital photography with flash and no-flash image pairs. In *ACM transactions on graphics (TOG)*.

E. Salvador, A. Cavallaro, and T. Ebrahimi. 2001. Shadow identification and classification using invariant color models. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Vol. 3. 1545–1548 vol.3.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. 2019. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *CVPR*.

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum. 2006. Flash Matting. *ACM Trans. Graph.* 25, 3 (jul 2006), 772–778.

Jian Sun, Jian Sun, Sing Bing Kang, Zong-Ben Xu, Xiaoou Tang, and Heung-Yeung Shum. 2007. Flash Cut: Foreground Extraction with Flash and No-flash Image Pairs. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

A. Yuille and D. Snow. 1997. Shape and albedo from multiple images using integrability. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 158–164.

Changyin Zhou, Alejandro Troccoli, and Kari Pulli. 2012. Robust stereo with flash and no-flash image pairs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 342–349.