UNIVERSITY OF CALIFORNIA

Los Angeles

Combining Physics with Machine Learning:
Case Study of Shape from Polarization

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Electrical and Computer Engineering

by

Yunhao Ba

2019

ABSTRACT OF THE THESIS

Combining Physics with Machine Learning:
Case Study of Shape from Polarization

by

Yunhao Ba
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2019
Professor Achuta Kadambi, Chair

Shape from Polarization (SfP) recovers an object's shape from polarized photographs of the scene. In previous works, the SfP algorithms use idealized physical equations to recover the shape. These previous approaches are error-prone when real-world conditions deviate from the idealized physics. In this thesis, we propose a physics-based neural network to address the SfP problem. Our algorithm fuses deep learning with synthetic renderings (derived from physics) to exceed the quality of all previous SfP methods. A two-stage encoder is used to resolve the longstanding problem of ambiguities. Our results of surface normal recovery are an improvement upon methods that utilize physics-based solutions alone.

The thesis of Yunhao Ba is approved.

Vwani P. Roychowdhury

Subramanian Srikantes Iyer

Achuta Kadambi, Committee Chair

University of California, Los Angeles

2019

To my beloved parents and family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This thesis is an accumulative work of all my research experiences, and I would like to express my sincere gratitude to so many people who have helped and supported me.

First, I sincerely appreciate my advisor, Prof. Achuta Kadambi for his detailed guidance and unconditional support in academics. This thesis could not become possible without his supervision and suggestions. I am also grateful to my thesis committee members: Prof. Subramanian S. Iyer and Prof. Vwani P. Roychowdhury for their knowledge and kind support for this thesis. Further, I want to thank Prof. Wan-Chi Siu, my advisor at the Hong Kong Polytechnic University, who offered me research opportunities during undergraduate studies.

I would like to acknowledge all the co-authors for the collaborative work within this thesis. Prof. Boxin Shi from Peking University deserves a special thanks for his insightful ideas and timely assistance, and I am grateful to Yiqin Wang, Rui Chen, and Prof. Lei Yan for their hard work on data preparation. I also truly thank NVIDIA's Academic Programs Team for their GPU support.

Finally, please allow me to express my deepest gratitude to my parents and all other family members. Their emotional support and encouragement are indispensable along my journey of research.

VITA

2013–2017	BEng(Hons) in Electronic and Information Engineering Department of Electronic and Information Engineering The Hong Kong Polytechnic University
2017–2019	MS in Electrical and Computer Engineering Department of Electrical and Computer Engineering University of California, Los Angeles (UCLA)

CHAPTER 1

Introduction

1.1 Overview

This thesis studies how physical priors can be combined with neural networks, using a difficult imaging problem as a case study. There is a logic to combining model-based priors with neural networks. Model-based methods may robustly describe simple parts of the system, while neural networks can describe complex portions of a system, where closed-form models may not be available. Additionally, as compared to using naive neural networks alone, the use of models enables a form of regularization on the neural network output. The end-result, as we show in this thesis, is a performance improvement and a potential step toward guaranteeing physically realizable solution spaces.

Combining physics and deep learning is a problem not without subtleties. As we study in this thesis, many of the physical models are centuries old - how do we create a common embedding space where neural network weights map to physical equations? These questions are central, not only to this thesis, but also recentw work. For example, [KWR17, PLD18, AWF18] show the effectiveness of combining physical knowledge with deep learning, while [DSH17, SSO18, LYC17] exhibit the advantages of incorporating deep learning with some existing models. However, there is a key differentiator between prior work and this thesis. Most of the existing methods assume that the physical models that characterize the system are known. This thesis scales to the new goal of when there is considerable uncertainty in physical equations as well as pattern recognition. For that reason, this thesis selects an inverse problem in imaging that is known to have a poor physical solution: Shape from Polarization (SfP).

We have picked this problem as an ideal case study because its physical solutions are formulated under ideal assumptions, and suffer from the ambiguity problem, and the polarization data is usually noisy, which is difficult for normal deep learning frameworks to learn from. While we are advancing our understanding of the broad question of physics-based machine learning, we are also making progress toward the concrete computer vision problem of SfP.

1.2 Shape from Polarization

Inverse problem is the process of retrieving system parameters based on the observations of the system. The general difficulties of inverse problems might be summarized as follows: **1.** Observations may not be sufficient to recover the whole system; **2.** Obtaining the exact parameters usually requires search within a massive parameter space. SfP is a classic inverse problem in physics-based computer vision. The goal is to recover the shape of an object from polarized photographs of a scene. The motivation is easy to grasp: light reflecting off an object has a polarization state that relates to the object's shape. Unfortunately, recovering surface normals through SfP is very difficult.

One physical challenge in SfP is the *ambiguity problem.* This problem arises because a linear polarizer cannot distinguish between polarized light that is rotated by π radians. This results in two confounding estimates for the azimuth angle. Previous work in SfP has used additional information to constrain the ambiguity problem. For instance, [MEF12] use shape from shading constraints to correct the ambiguities. Other authors assume surface convexity to constrain the azimuth angle [MTH03, AH06]. Yet another solution is to use a coarse depth map to constrain the ambiguity [KTS15, KTS17]. Figure 2.2 compares the tradeoffs of our proposed technique with these alternatives.

Another physical challenge in SfP is the *refractive problem.* SfP requires knowledge of per-pixel refractive indices. Previous work has used hard-coded values to estimate the refractive index of scenes. This leads to a relative shape that is recovered with refractive distortion. Another physical challenge is the *noise problem.* SfP is ill-conditioned, requiring

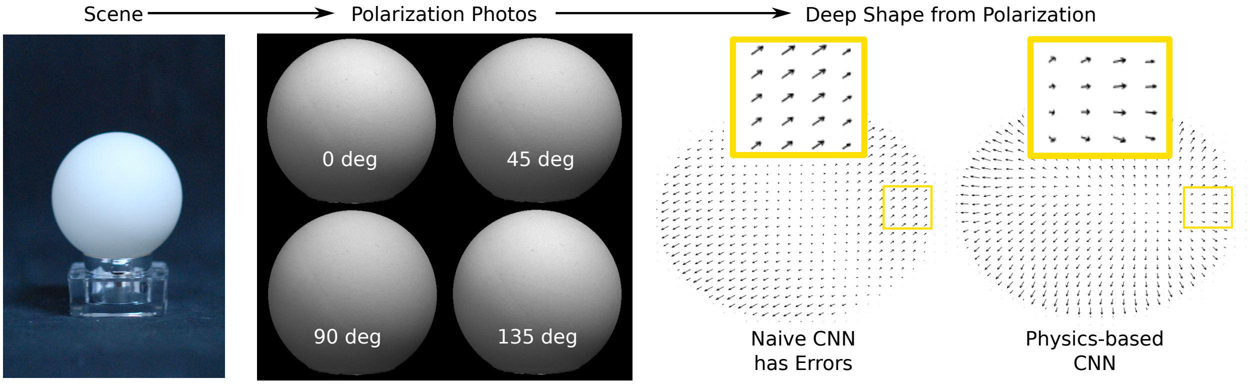


Figure 1.1: **The novelty of this thesis is to use physics-based neural networks to address the shape from polarization (SfP) problem.** Ordinarily, SfP uses physics-based equations to convert polarized photos into surface normal maps. Here, we show two machine learning approaches for SfP. Using learning-only results in artifacts on a simple object, while a physics-based neural network is more successful. Please see the inset for comparison.

input images that are relatively noise-free. Ironically, a polarizer reduces the captured light intensity by 50 percent, worsening the effects of Poisson shot noise.

We address these SfP pitfalls by moving away from a physics-only solution, toward the realm of data-driven techniques. A reasonable first attempt could apply vanilla convolutional neural networks (CNN) to the SfP problem. Unfortunately, machine learning alone is not a satisfactory solution. As illustrated in Figure 1.1, a naive CNN implementation does not work even on the simplest of scenes. In contrast to prior work, we fuse both physics and deep learning in symbiosis. This hybrid approach outperforms previous SfP methods.

1.3 Contributions

In context of prior work in SfP, this thesis demonstrates three technical first attempts:

1. implementing deep learning for the SfP problem;
2. incorporating physics into the deep learning approach; and
3. providing the first dataset to enable data-driven SfP.

Scope: Because this is only a first attempt, the proposed solution is not perfect, particularly when obtaining the shape of objects with mixed reflectivities. However, all prior methods in SfP fail in this scenario. While our physics-based approach to neural networks does outperform the individual strategy of physics and learning alone, this may just be a first attempt at the problem.

1.4 Notation

Symbol	Meaning
ϕ	phase of received light sinusoid
I_{max}	maximum intensity of of received light sinusoid
I_{min}	minimum intensity of of received light sinusoid
ϕ_{pol}	polarizer rotation angle
$I(\phi_{pol})$	intensity of received light sinusoid with a polarizer angle of ϕ_{pol}
$\boldsymbol{I}_{\phi_{pol}}$	polarized image with a polarizer angle of ϕ_{pol}
φ	azimuth angle
θ	zenith angle
ρ	degree of polarization
n	refractive index
W	width of the image
H	height of the image
$\boldsymbol{N}$	ground truth surface normal map
$\hat{\boldsymbol{N}}$	estimated surface normal map from the network
$\boldsymbol{N}_{diff}$	surface normal map from diffuse polarization model
$\boldsymbol{N}_{spec1}$	the first surface normal map from specular polarization model
$\boldsymbol{N}_{spec2}$	the second surface normal map from specular polarization model

Table 1.1: **Notation table**

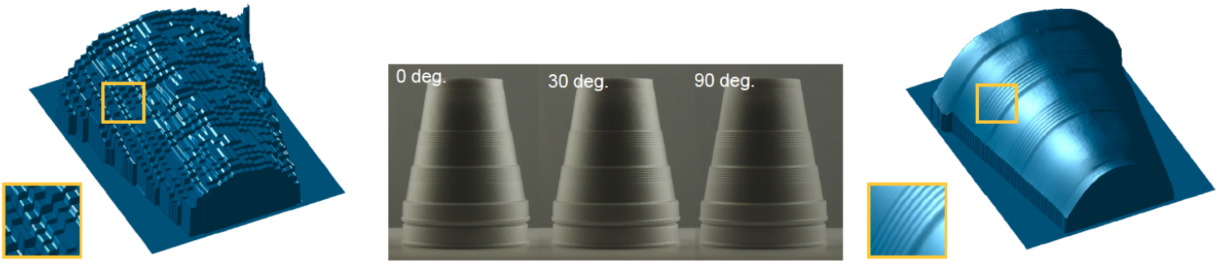
CHAPTER 2

Background

2.1 Related Work

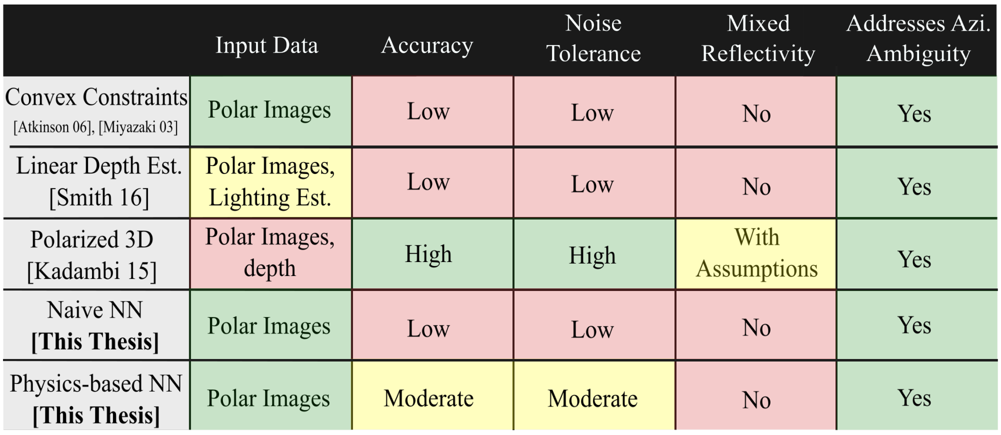
Polarized light has exhibited remarkable potentials on shape recovery as shown in Figure 2.1. However, there is no previous work that takes the advantage of data-driven deep learning techniques to enhance the performance. This thesis sits at the seamline of deep learning and SfP, offering unique performance tradeoffs from prior work. Refer to Figure 2.2 for an overview.

Shape from polarization infers the shape (usually represented in surface normal) of a surface by observing the correlated changes of image intensity with the polarization information. Changes of polarization information could be captured by rotating a linear polarizer in front of an ordinary camera [Wol97, AE18] or polarization cameras using a single shot in real time (e.g., PolarM [Pol17] camera used in [YTL18]). Conventional shape from polarization decodes such information to recover the surface normal up to some ambiguity. If only images with different polarization information are available, heuristic priors such as the surface normals on the boundary and convexity of the objects are employed to remove the ambiguity [MTH03, AH06]. Photometric constraints from shape from shading [MEF12] and photometric stereo [Dv01, NNT15, Atk17] complements polarization constraints to make the normal estimates unique. If multi-spectral measurements are available, surface normal and its refractive index could be estimated at the same time [HRH10, HRH13]. More recently, a joint formulation of shape from shading and shape from polarization in a linear manner is shown to be able to directly estimate the depth of the surface [SRT16, TSZ17, SRT18]. This thesis is a first attempt at studying deep learning and SfP together.



(a) Depth from Microsoft Kinect **(b)** Three Photos using a Polarizer **(c)** Polarization Enhanced Depth

Figure 2.1: **The power of polarized light for shape recovery** (a) The Kinect depth of a cup (b) Three polarized images with different polarizer angles (c) The enhanced depth with polarization cues. (This figure is adapted from [KTS15])



	Input Data	Accuracy	Noise Tolerance	Mixed Reflectivity	Addresses Azi. Ambiguity
Convex Constraints [Atkinson 06], [Miyazaki 03]	Polar Images	Low	Low	No	Yes
Linear Depth Est. [Smith 16]	Polar Images, Lighting Est.	Low	Low	No	Yes
Polarized 3D [Kadambi 15]	Polar Images, depth	High	High	With Assumptions	Yes
Naive NN **[This Thesis]**	Polar Images	Low	Low	No	Yes
Physics-based NN **[This Thesis]**	Polar Images	Moderate	Moderate	No	Yes

Figure 2.2: **Summarizing the tradeoffs of our proposed physics-based neural networks (NN) versus physics-only and learning-only approaches.**

Polarized 3D involves stronger assumptions than SfP and has different inputs and outputs. Recognizing that SfP alone is a limited technique, the Polarized 3D class of methods integrate shape from polarization with a low resolution depth estimate. This additional constraint allows not just recovery of shape but also a high-quality 3D model. The low resolution depth could be achieved by employing two-view [MKI04, AH05, BVM17], three-view [CZS18], multi-view [MSB16, CGS17] stereo, or even in real time by using a SLAM system [YTL18]. These depth estimates from geometric methods are not reliable in textureless regions where finding correspondence for triangulation is difficult. Polarimetric cues could be jointly used to improve such unreliable depth estimates to obtain a more complete shape estimation. A depth sensor such as the Kinect can also provide coarse depth prior to disambiguate the ambiguous normal estimates given by SfP [KTS15, KTS17]. The key step that characterizes Polarized 3D is a holistic approach that rethinks both SfP and the depth-normal fusion process. The main limitation of Polarized 3D is the strong requirement that there be a coarse depth map, which is not true for our proposed technique.

Data-driven computational 3D imaging approaches draw much attention in recent years thanks to the powerful modeling ability of deep neural networks. Various types of convolutional neural networks (CNNs) are designed and trained to enable 3D imaging for different types of sensors and measurements. From single photon sensor measurements, a multi-scale denoising and upsampling CNN is proposed to refine depth estimates [LOW18]. CNNs also show advantage in solving phase unwrapping, multipath interference, and denoising jointly from the raw time-of-flight measurements [MHM17, SHW18]. From multi-directional lighting measurements, a fully-connected network is first proposed to solve photometric stereo for general reflectance with a pre-defined set of light directions [SSS17]. Then the fully-convolutional network with an order-agnostic max-pooling operation [CHW18] and the observation map invariant to the number and permutation of the images [Ike18] are concurrently proposed to deal with an arbitrary set of light directions. Normal estimates from photometric stereo can also be learned in an unsupervised manner by minimizing the reconstruction loss [TM18]. The challenge with existing deep learning frameworks is that they do not leverage the unique physics of polarization.

2.2 Physical Solution

Our objective is to reconstruct surface normals $\hat{\boldsymbol{N}}$ from a set of polarized images $\{\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}$, ..., $\boldsymbol{I}_{\phi_M}\}$ with different rotations of polarizer angles. For a specific polarizer angle ϕ_{pol}, the intensity at a pixel of a captured image follows a sinusoid variation under unpolarized illumination:

$$I(\phi_{pol}) = \frac{I_{max} + I_{min}}{2} + \frac{I_{max} - I_{min}}{2} \cos(2(\phi_{pol} - \phi)), \tag{2.1}$$

where ϕ denotes the phase angle, and I_{min} and I_{max} are lower and upper bounds for the observed intensity. Equation 2.1 has a π-**ambiguity** in context of ϕ: two phase angles, with a π shift, will result in the same intensity in the captured images. Based on the phase angle ϕ, the azimuth angle φ can be retrieved with $\frac{\pi}{2}$-**ambiguity** as follows [CGS17]:

$$\phi = \begin{cases} \varphi, & \text{if diffuse reflection dominates} \\ \varphi - \frac{\pi}{2}, & \text{if specular reflection dominates} \end{cases}. \tag{2.2}$$

The zenith angle θ is related to the degree of polarization ρ, which can be written as:

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}. \tag{2.3}$$

When diffuse reflection is dominant, the degree of polarization can be expressed with the zenith angle θ and the refractive index n as follows [AH06]:

$$\rho = \frac{(n - \frac{1}{n})^2 \sin^2\theta}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2\theta + 4\cos\theta\sqrt{n^2 - \sin^2\theta}}. \tag{2.4}$$

The effect of n is not decisive, and we assume $n = 1.5$ throughout the rest of this thesis. With this known n, Equation 2.4 can be rearranged to obtain a close-form estimation of the zenith angle for the diffuse dominant case.

When specular reflection is dominant, the degree of polarization can be written as [AH06]:

$$\rho = \frac{2\sin^2\theta\cos\theta\sqrt{n^2 - \sin^2\theta}}{n^2 - \sin^2\theta - n^2\sin^2\theta + 2\sin^4\theta}. \tag{2.5}$$

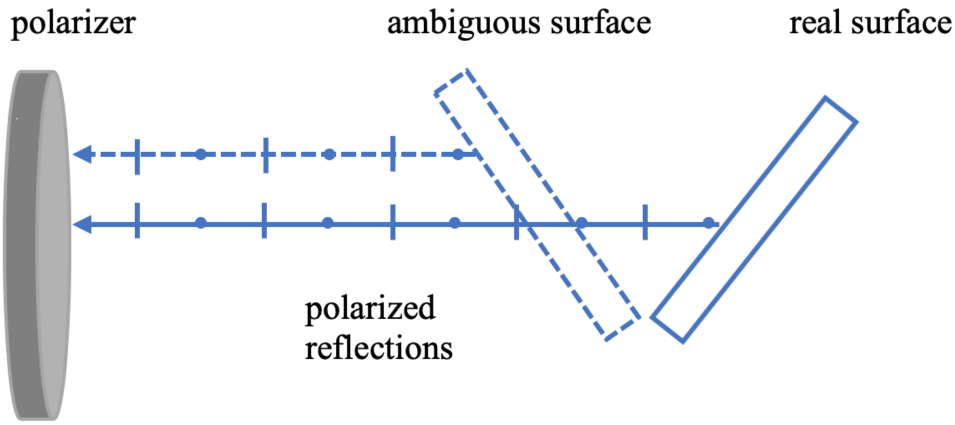


Figure 2.3: **SfP lacks a unique solution due to the *ambiguity problem*.** Here, two different surface orientations could result in the same exact polarization signal, represented by dots and hashes. The dots represent polarization out of the plane of the thesis and the hashes represent polarization within the plane of the board. Based on the measured data, it is unclear which orientation is correct.

Equation 2.5 can not be inverted analytically, and solving the zenith angle with numerical interpolation will produce two solutions if we do not introduce additional constraints. For real world objects, specular reflection and diffuse reflection are mixed depending on the surface material of the object. As shown in Figure 2.3, the ambiguity in the azimuth angle and uncertainty in the zenith angle are fundamental limitations of SfP. Overcoming these limitations through physics-based neural networks is the primary focus of this thesis.

CHAPTER 3

Physics-based Learning Model

3.1 Learning with Physics

Large amounts of labeled data are critical to the success of neural networks. To alleviate the burden of data requirement, one possible method is to incorporate physical priors during learning. However, it is essentially difficult to use physical information for SfP tasks due to the following reasons: **1.** Polarization normals contain ambiguous azimuth angles. **2.** Specular reflection and diffuse reflection coexist simultaneously, and determining the proportion of each type is complicated. **3.** Polarization normals are usually noisy, especially when the degree of polarization is low. Shifting the azimuth angles by π or $\frac{\pi}{2}$ could not reconstruct the surface normals properly for noisy images.

Due to the above reasons, regularization from the physical azimuth angle or the physical zenith angle will degrade the network performance and lead to a fragile model. Therefore, instead of using physical solutions as regularization, we directly feed both the polarized images and the ambiguous surface normals into the network, and leave the network to learn how to combine physical solutions with the polarized images effectively. The estimated surface normals can be structured as following:

$$\hat{\boldsymbol{N}} = f(\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, ..., \boldsymbol{I}_{\phi_M}, \boldsymbol{N}_{diff}, \boldsymbol{N}_{spec1}, \boldsymbol{N}_{spec2}) \tag{3.1}$$

where $f(\cdot)$ is the proposed prediction model, $\{\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, ..., \boldsymbol{I}_{\phi_M}\}$ is a set of polarized images, and $\hat{\boldsymbol{N}}$ is the estimated surface normals. We use diffuse model in Section 2.2 to calculate $\boldsymbol{N}_{diff}$, and $\boldsymbol{N}_{spec1}, \boldsymbol{N}_{spec2}$ are the two solutions from specular model.

The remaining question is to contrive a way to combine ambiguous surface normals with polarized images in the network. Simply concatenating $\boldsymbol{N}_{diff}, \boldsymbol{N}_{spec1}, \boldsymbol{N}_{spec2}$ with

Layer	Encoder block
1	Conv[(3×3), m, $m \times 2$, stride=2], BN, LeakyReLU
2	Conv[(3×3), $m \times 2$, $m \times 2$, stride=1], BN, LeakyReLU
3	Conv[(3×3), $m \times 2$, $m \times 2$, stride=1], BN, LeakyReLU
Layer	**Decoder block**
1	Deconv[(3×3), m, $\frac{m}{2}$, stride=2], BN, LeakyReLU
2	Conv[(3×3), $\frac{m}{2}$, $\frac{m}{2}$, stride=1], BN, LeakyReLU
3	Conv[(3×3), $\frac{m}{2}$, $\frac{m}{2}$, stride=1], BN, LeakyReLU

Table 3.1: **Convolutional layers in each encoder-block and decoder-block.** Conv[$(k \times k)$, a, b, stride=s] represents a 2D convolutional layer with kernel size of $(k \times k)$, a input channels, b output channels, and s stride. Deconv denotes a 2D transposed convolutional layer, and BN denotes a batch normalization layer. We use LeakyReLU [MHN13] with a negative slope of 0.1 as the activation function.

polarized images did not show us the expected enhancement based on our testing results. One explanation for that is the low-level features from polarized images and the low-level features from ambiguous normals are different, and it is burdensome for convolutional layers to learn these two types of features concurrently. An alternative method is to use two separate encoder streams to encode these two types of features at the low-level stage, and merge the high-level features in deeper layers. With the proposed two-stream encoder, ambiguous normals can implicitly direct the network to learn some physical information and serve as a good initialization to improve generalizability.

3.2 Network Architecture

Our network structure is illustrated in Figure 3.1. It consists of two independent encoders to extract features from polarized images and ambiguous surface normals separately and a common decoder to output surface normal $\hat{\boldsymbol{N}}$. A variation of U-Net [RFB15] and LinkNet [CC17]

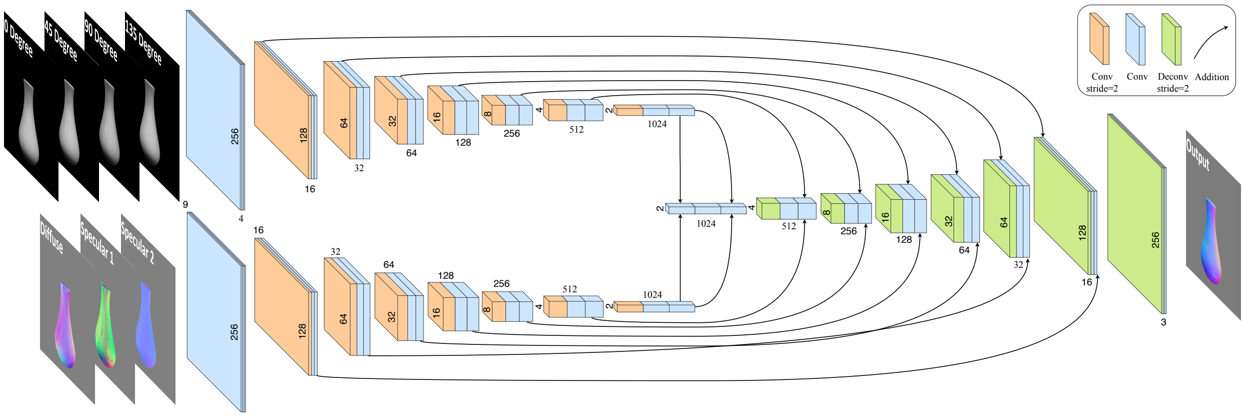


Figure 3.1: **Overview of our proposed physics-based neural networks.** The network is designed according to the encoder-decoder architecture in a fully convolutional manner. We use addition operation as the mixer to integrate both low-level and high-level features from polarized images and ambiguous surface normals.

is used to connect encoder block and decoder block at the same hierarchical level. We argue that addition is superior to concatenation when merging feature maps, since it achieves comparable performance, yet requires less memory and computational power in general based on our testing results.

There are 7 encoder blocks to encode the input into a tensor of dimensionality $B \times 1024 \times 2 \times 2$ to guarantee the receptive field, where B is the minibatch size. The encoded tensor is then decoded by the same number of decoder blocks to produce the estimated surface normals $\hat{\boldsymbol{N}}$. An ℓ_2-normalization layer is appended after the last decoder block to convert corresponding feature maps into surface normals. Table 3.1 shows the structure of each encoder and decoder block. Two additional feature extractors containing 3 convolutional layer of kernel size 3×3 are placed before the first encoder block to prepare feature maps suitable for downsampling purpose. We use convolutional layers with stride of 2 for downsampling, and transposed convolutional layers for upsampling. Batch normalization layers [IS15] are inserted after each layer, except the output layer, where batch normalization would cause distortion of the estimated surface normals $\hat{\boldsymbol{N}}$. After batch normalization, LeakyReLU with a negative slope of 0.1 is used for the activation function.

For the image encoder, pictures captured with a polarizer at angles $\phi_{pol} \in \{0°, 45°, 90°, 135°\}$ are selected for training and testing. It is sufficient to solve the polarization cues with three values of ϕ_{pol}, nevertheless we use four values to ensure the robustness over noise. The four polarized images are stacked to form a tensor of dimensionality $4 \times H \times W$, where $H \times W$ is the spatial resolution of polarized images. Our motivation is that, since the relative 3D information from polarization is essentially from the the intensity difference between polarized images, it is beneficial for convolutional layers to learn this difference by concatenating images along the channel dimension as input. For the normal encoder, we use the identical architecture for the sake of feature map addition. We use ground truth surface normals to supervise the physics-based neural networks with the cosine similarity loss function:

$$L_{cosine} = \frac{1}{W \times H} \sum_{i}^{W} \sum_{J}^{H} (1 - \langle \hat{\boldsymbol{N}}_{ij}, \boldsymbol{N}_{ij} \rangle), \tag{3.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, $\hat{\boldsymbol{N}}_{ij}$ is the estimated surface normal at pixel location (i, j), and $\boldsymbol{N}_{ij}$ is the corresponding ground truth of surface normal. This loss is minimized when $\hat{\boldsymbol{N}}_{ij}$ and $\boldsymbol{N}_{ij}$ have identical orientation.

CHAPTER 4

Dataset and Implementation Details

4.1 Dataset

To train the physics-based neural network, polarization images with corresponding normal maps are needed. However, neither synthetic nor real datasets for such a purpose are publicly available. We therefore create the first real and synthetic datasets for data-driven SfP as shown in Figure 4.2.

Real dataset: A camera with a layer of polarizers above the photodiodes [Luc18] is used to capture four polarized images at angles $0°, 45°, 90°$ and $135°$ in a single shot. Then a structured light based 3D scanner [SHI18] (with single shot accuracy no more than 0.1 mm, point distance from 0.17 mm to 0.2 mm, and a synchronized turntable for automatically registering scanning from multiple viewpoints) is used to obtain high-quality 3D shapes. Our real data capture setup is shown in Figure 4.1. The scanned 3D shapes are aligned from the scanner's coordinate system to the image coordinate system of the polarization camera by using the shape-to-image alignment method adopted in [SMW19]. Finally, we compute for surface normals of the aligned shapes by using the Mitsuba renderer [Jak10] as ground truth. In total, we capture 65 sets (with 4 polarized images plus a surface normal map) of real data, and we use 58 sets of them for training and the remaining 7 sets for testing and quantitative evaluation.

Synthetic dataset: The scanned real data are not sufficient in terms of scale and lighting variation for training a deep neural network. We further create a synthetic dataset to complement the real one. We use the normal maps provided in [SMW19], since they cover a great

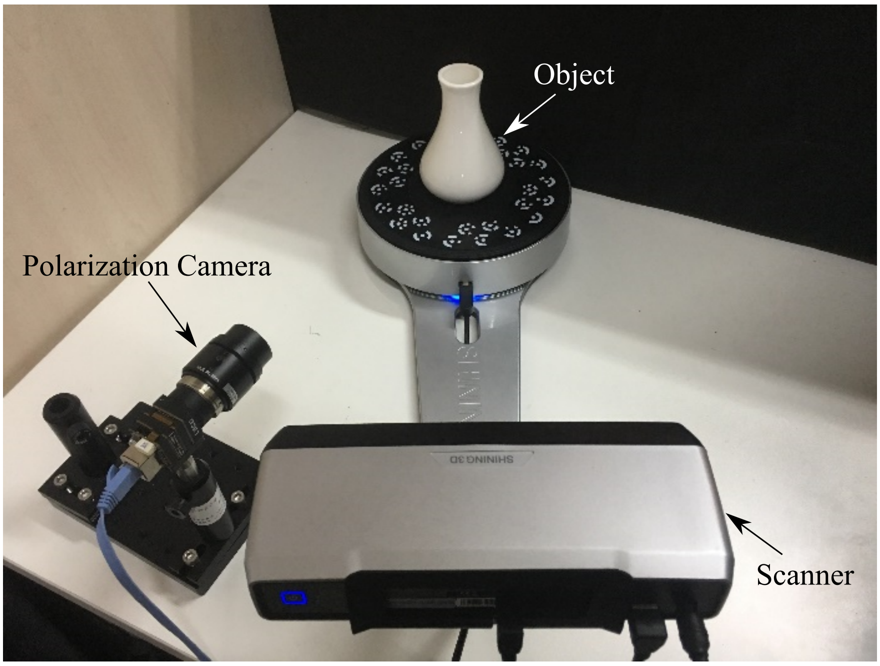


Figure 4.1: **Physical setup to capture polarized images.** We use a polarization camera to capture four polarimetric measurements of an object in a single shot. The scanner is put next to the camera for obtaining the 3D shape of the object. The setup is put in an indoor environment with typical office lighting.

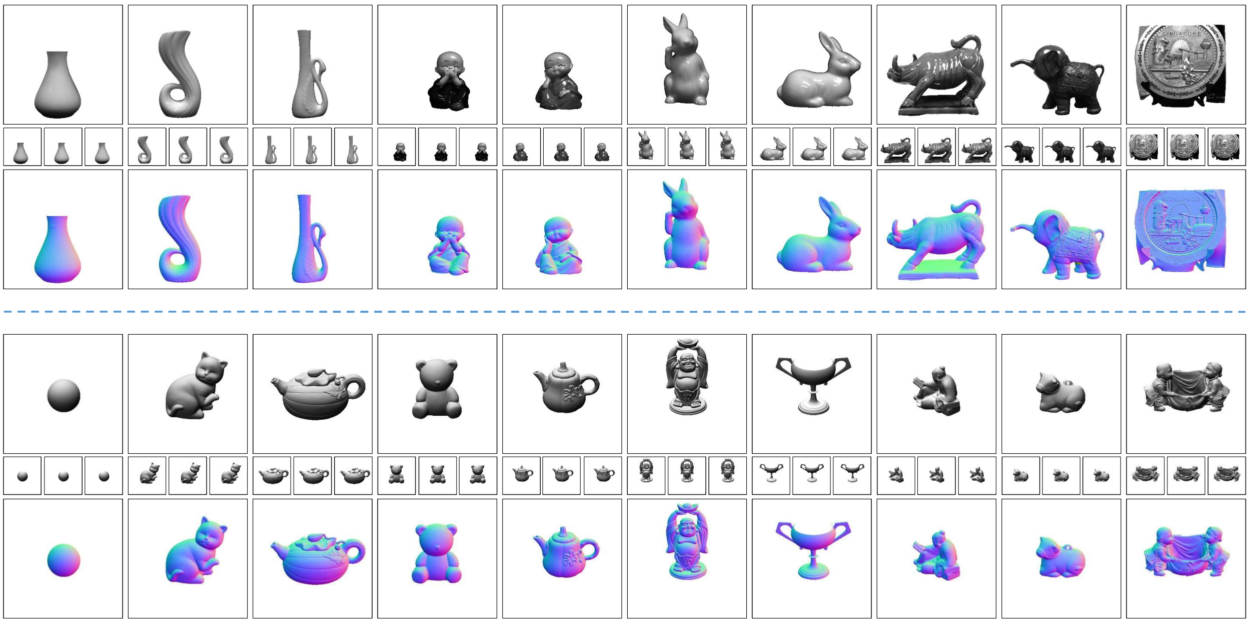


Figure 4.2: **Overview of our real (upper part) and synthetic (lower part) datasets.** We show 10 objects (out of 58) in the training set of our real dataset, and 10 objects (out of 10) of our synthetic dataset. In each example, we show $\boldsymbol{I}_0$ on top of $\boldsymbol{I}_{45}, \boldsymbol{I}_{90}, \boldsymbol{I}_{135}$ with thumbnail sizes, and the corresponding normal maps are shown below the polarization images. Note the polarization camera captures gray scale images, which are used as input for computation.

diversity of geometry from a simple sphere to surfaces with highly delicate structures. Given a normal map, we calculate the diffuse shading by assuming the Lambertian reflectance and a distant environment map [Deb08], as $\boldsymbol{I}_0$, $\boldsymbol{I}_{45}, \boldsymbol{I}_{90}, \boldsymbol{I}_{135}$ are calculated using Equation 2.1. By using 10 different environment maps on 10 different normal maps, we obtain 100 sets of synthetic data, and all these data are used for training.

4.2 Software Implementation

Our model was implemented in PyTorch [PGC17], and trained for 500 epochs with a batch size of 64. It took 8 hours for the network to converge with a single NVIDIA Titan V GPU. We used Adam optimizer [KB14] with default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$), and the base learning rate was set to be 0.01. The learning rate was multiplied with a factor of 0.8 when loss reached the plateau regions during the training process. We tried both He initialization [HZR15] and Xavier [GB10] on the convolutional weights, and the performance of Xavier initialization is slightly better. For data augmentation, images patches of size 256 × 256 are randomly cropped during training, and a patch is discarded if its foreground ratio is less than 20%. No random rescaling is used to preserve the original high-resolution details and aspect ratio. The final prediction is the average of 32 shifted input to preserve the accuracy at boundaries of each patch.

4.3 Implementing Comparisons to Physics-only SfP

We used a test dataset consisting of scenes that include BALL, HORSE, VASE, HALF PAINTED VASE, CHRISTMAS, FLAMINGO, RABBIT. On this test set, we compared performance between our proposed method and three physics-only methods for SfP: **1.** [SRT16]. **2.** [MEF12]. **3.** [MTH03, AH06]. The first method recovers the depth map directly, and we only use the diffuse model due to the lack of specular reflection masks. The surface normals are obtained from the estimated depth with bicubic fit. Both the first and the second methods require lighting input, and we use the estimated lighting from the first method during comparison. The second method also requires known albedo, and following convention, we assume a constant albedo. All the comparison codes were provided by Smith et al. [SRT16] [1].

[1] https://github.com/waps101/depth-from-polarisation

CHAPTER 5

Result and Discussion

In this section, we evaluate our model with the presented challenging real-world scene benchmark, and compare it against three physics-only methods for SfP. Mean angular error (MAE) is selected as the metric to quantify the accuracy of the estimated surface normals during comparison.

5.1 Machine Learning Alone is Insufficient

As illustrated in Figure 1.1, a naive approach to deep learning that does not incorporate physics is insufficient. On one of the simplest scenes possible (a white ping-pong ball), the naive neural network cannot recover accurate surface normals. There is only slight difference between images with different polarized angles, and it is difficult for a naive neural net to learn from these differences with limited number of training samples. The proposed method incorporates multiple physical solutions. Therefore, apart from learning from pure polarized images, which is difficult, the network can also learn from physical solutions, which may be easier. Generalizability of the network is thus improved, and it becomes realistic for the network to predict high-quality normals in this case.

5.2 Choice of Loss Function is Important

As shown in Figure 5.1, the choice of loss function affects both the quantitative error and the recovery of qualitative detail. Use of the ℓ_2 loss function results in an overall smoothened result, while the ℓ_1 shows widening of the ridges in the vase. The cosine loss function is

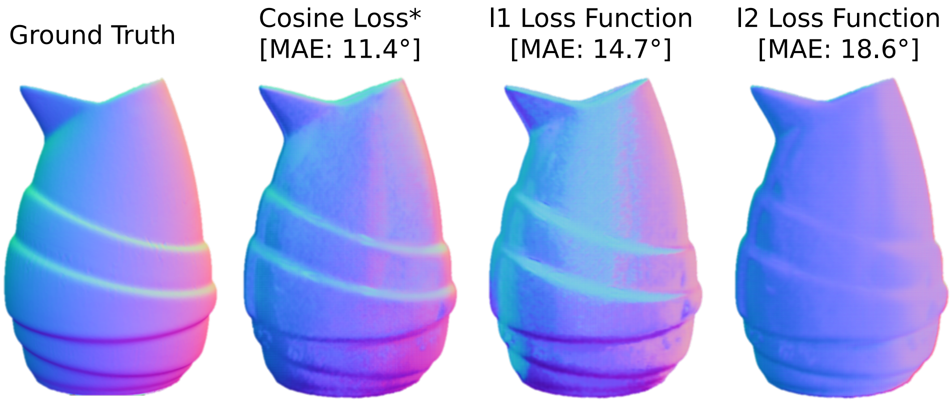


Figure 5.1: **Choice of neural network loss function affects result quality.** Motivated by this example, we choose the cosine loss function as it returns the lowest error and appears to recover relatively more detail. Compared results are obtained on a small training set with 32 training samples (16 synthetic samples, 16 real samples).

closest to the ground truth and is used in all other scenes from the thesis. The success of cosine loss may come from its emphasis on the orientation information. Both ℓ_1 and ℓ_2 loss will penalize the length of estimated surface normals, however, the normalization layer at the end has already constrained the normal length.

5.3 Improved Performance on Shiny and Detailed Scene

Here, we show improved performance on a relatively shiny scene with surface details. As illustrated in Figure 5.2, the proposed method of physics-based NN achieves the highest qualitative and quantitative accuracy. Worth noting is that, the result from [SRT16] does not perform well on the HORSE SCENE because the simple hybrid reflection model and spherical harmonics based lighting model are not well satisfied for HORSE SCENE, and the estimated depth becomes inaccurate, which results in a normal map with a large error.

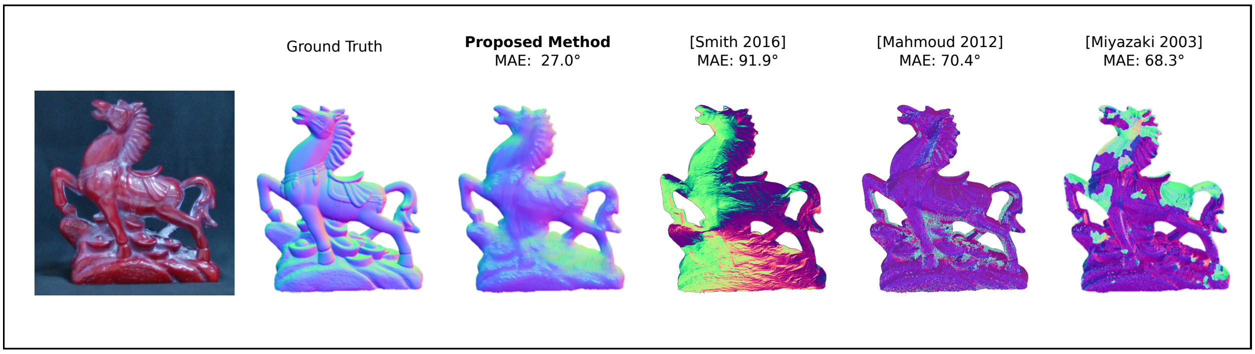


Figure 5.2: **Our method can handle shiny scenes with high-frequency details.** Although the proposed method does not recover all of the detail that was present in ground truth, global errors in shape are not present. By comparison, the physics-only methods exhibit large errors in shape recovery.

5.4 Improved Performance in Noise-degraded Environments

Here, we show that the physics-based NN approach outperforms physics-only approaches when the noise level drops. As illustrated in Figure 5.3, the input to each of the methods are noisy polarization images. This noise was generated in simulation to mimic low light levels (when shot noise dominates). The proposed physics-based NN approach shows a qualitative and quantitative improvement over the physics-only methods. Our proposed approach of using a physics-based neural network works in low noise levels because of the encoder-decoder architecture. Both polarized images and physical solutions will be downsampled into a condensed feature map by the encoder, and the decoder has to use this condensed feature map to recover the normal map. With limited number of parameters, the network has to learn some intrinsic representation of the input, which gives us the robustness over noise.

5.5 Additional Scenes

Over all tested scenes in the thesis, the proposed physics-based neural network outperforms physics-only methods from [MTH03, MEF12, SRT16]. In particular, Figure 5.4 shows that the proposed method recovers surface normals that are quantitatively and qualitatively clos-

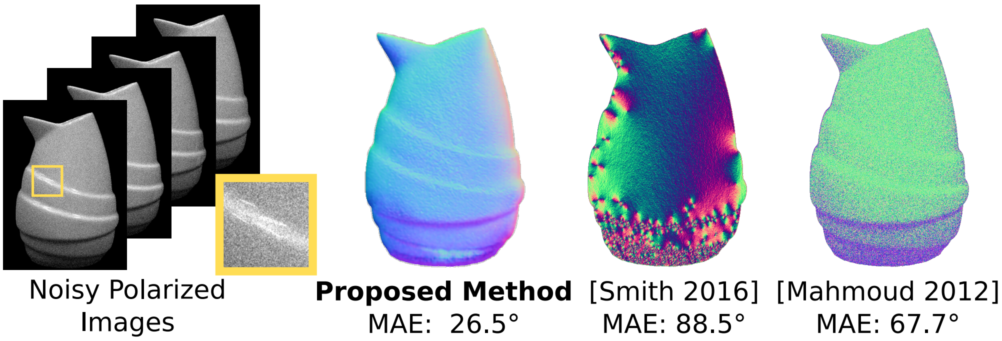


Figure 5.3: **The proposed method handles cases when the input images are noisy**. Noise-tolerant performance is particularly important when using polarizing optics; a polarizing filter reduces the light intensity by 50 percent.

est to ground truth. The large region-wise anomalies on many of the results from [MTH03] are to do with the region-growing constraint on the convexity that is imposed. The method of [MEF12] uses shading constraints which require a distant light source, which is not the case for tested scenes. Finally, the results in [SRT16] are explained both by the use of 4 polarized images as input (ordinarily the method requires 18), as well as change in the lighting direction.

5.6 SfP Still Fails on Mixed Material Scenes

This thesis, like other SfP methods, is unable to solve the *mixed material* problem. This problem occurs when the polarimetric signal is not just due to surface geometry, but also material effects. Figure 5.5 shows one such scene, consisting of a vase painted with two different styles of paint. While the physics-based NN result has the lowest quantitative error, none of the SfP methods are correct. There is a texture copy artifact at the point where the paints change.

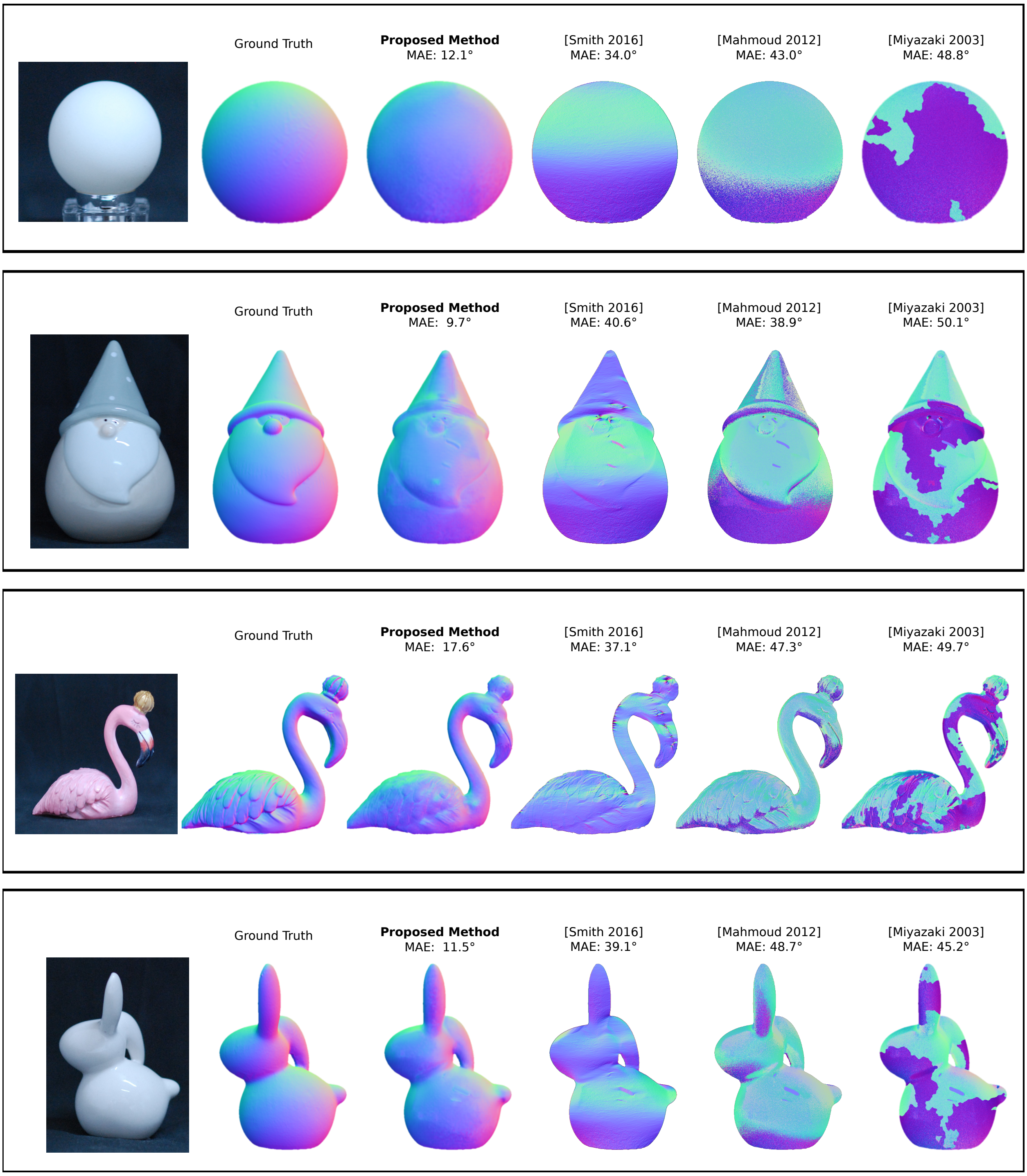


Figure 5.4: **The proposed method has the least error in recovering normal maps.** We compare with SfP papers from [SRT16], [MEF12] and [MTH03]. Not shown is the performance from [AH06], which behaves similarly to [MTH03].

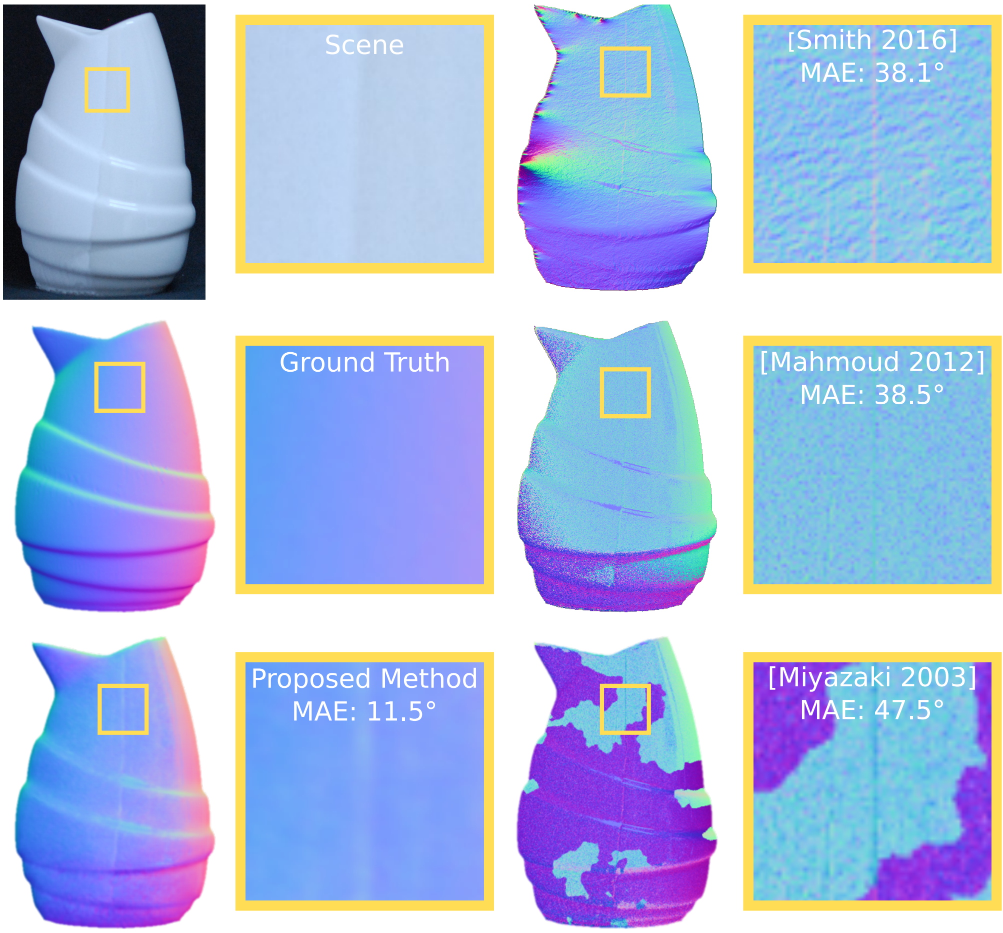


Figure 5.5: **All SfP methods, including the proposed method, fail on a scene with mixed paints.** A texture copy artifact is seen in all the SfP methods at the point of material transition. While all SfP methods can be seen as failing in that regard, the proposed method still has the lowest error.

CHAPTER 6

Conclusion

In summary, we have presented a first attempt at physics-based deep learning to handle the challenging SfP problem. The proposed method is shown to outperform previous methods that do not leverage machine learning. Surprisingly, using only machine learning is also sub-optimal, even for the most simplistic of scenes (cf. Figure 1.1). This underscores the importance of incorporating physics into the deep learning pipeline.

Although our performance improvement holds for all tested scenes, several open problems remain unsolved. We find that existing SfP methods (including this thesis) fail on scenes with mixed reflectivity. It would be interesting to study how material properties could be incorporated into the physics-based NN architecture. Part of the solution may also rely on expanding the training dataset, to include a wider variety of object materials and paints. For these types of computational photography problems, where the capture procedure is labor intensive, it is likely that dataset sizes will be small. This underscores the importance of including physical priors in the network model. With this inclusion, we were able to obtain results from a relatively small dataset size.

The lessons learned in this ”Deep Shape from Polarization” study may also apply to a future ”Deep Polarized 3D” study. The physics-only family of Polarized 3D techniques benefit from robust integration of surface normals with a depth prior. The state-of-the-art Polarized 3D integration has been performed with a simplistic matrix inversion [KTS15]. A physics-based NN approach might be able to learn this elementary function to potentially obtain state-of-the-art results.

Beyond error metrics, there are other benefits to using deep learning pipelines. The feedforward pass can be computed in real-time. This is in contrast to physics-only SfP techniques, which tend to have long compute times [KTS15]. This benefit will become more apparent with the advancement of specialized computational architectures tailored toward deep learning. Overall, this thesis's results appear to validate the direction of jointly studying deep learning and SfP. The compassion between the naive neural network and the proposed physics-based neural network also verifies the effectiveness of physical information in the conventional deep learning architectures.

REFERENCES

[AE18] Gary A. Atkinson and Jürgen D. Ernst. "High-sensitivity analysis of polarization by surface reflection." *Machine Vision and Applications*, **29**(7):1171–1189, 2018.

[AH05] Gary A. Atkinson and Edwin R. Hancock. "Multi-view Surface Reconstruction using Polarization." 2005.

[AH06] Gary A Atkinson and Edwin R Hancock. "Recovery of surface orientation from diffuse polarization." **15**(6):1653–1664, 2006.

[Atk17] Gary A. Atkinson. "Polarisation photometric stereo." *Computer Vision and Image Understanding*, **160**:158–167, 2017.

[AWF18] Anurag Ajay, Jiajun Wu, Nima Fazeli, Maria Bauza, Leslie P Kaelbling, Joshua B Tenenbaum, and Alberto Rodriguez. "Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing." In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3066–3073. IEEE, 2018.

[BVM17] Kai Berger, Randolph Voorhies, and Larry H. Matthies. "Depth from stereo polarization in specular scenes for urban robotics." In *Proc. of International Conference on Robotics and Automation*, 2017.

[CC17] Abhishek Chaurasia and Eugenio Culurciello. "LinkNet: Exploiting encoder representations for efficient semantic segmentation." In *Proc. of IEEE International Conference on Visual Communications and Image Processing*, 2017.

[CGS17] Zhaopeng Cui, Jinwu Gu, Boxin Shi, Ping Tan, and Jan Kautz. "Polarimetric Multi-View Stereo." 2017.

[CHW18] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. "PS-FCN: A flexible learning framework for photometric stereo." 2018.

[CZS18] Lixiong Chen, Yinqiang Zheng, Art Subpa-asa, and Imari Sato. "Polarimetric Three-View Geometry." 2018.

[Deb08] Paul Debevec. "Rendering Synthetic Objects into Real Scenes: Bridging Traditional and Image-based Graphics with Global Illumination and High Dynamic Range Photography." In *ACM SIGGRAPH 2008 Classes*, pp. 32:1–32:10, 2008.

[DSH17] Steven Diamond, Vincent Sitzmann, Felix Heide, and Gordon Wetzstein. "Unrolled optimization with deep priors." *arXiv preprint arXiv:1705.08041*, 2017.

[Dv01] O. Drbohlav and R. Šára. "Unambiguous determination of shape from photometric stereo with unknown light sources." 2001.

[GB10] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2010.

[HRH10] Cong Phuoc Huynh, A. Robles-Kelly, and Edwin R. Hancock. "Shape and refractive index recovery from single-view polarisation images." 2010.

[HRH13] Cong Phuoc Huynh, A. Robles-Kelly, and Edwin R. Hancock. "Shape and refractive index from single-view spectro-polarimetric images." **101**(1):64, 2013.

[HZR15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." 2015.

[Ike18] Satoshi Ikehata. "CNN-PS: CNN-based photometric stereo for general non-convex surfaces." 2018.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167*, 2015.

[Jak10] Wenzel Jakob. "Mitsuba renderer.", 2010. http://www.mitsuba-renderer.org.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[KTS15] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. "Polarized 3d: High-quality depth sensing with polarization cues." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3370–3378, 2015.

[KTS17] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. "Depth sensing using geometrically constrained polarization normals." **125**(1-3):34–51, 2017.

[KWR17] Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. "Physics-guided neural networks (pgnn): An application in lake temperature modeling." *arXiv preprint arXiv:1710.11431*, 2017.

[LOW18] David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. "Single-Photon 3D Imaging with Deep Sensor Fusion." **37**(4):113, 2018.

[Luc18] Lucid Vision Phoenix polarization camera. "https://thinklucid.com/product/phoenix-5-0-mp-polarized-model/." 2018.

[LYC17] Hoang M Le, Yisong Yue, Peter Carr, and Patrick Lucey. "Coordinated multi-agent imitation learning." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1995–2003. JMLR. org, 2017.

[MEF12] Ali H Mahmoud, Moumen T El-Melegy, and Aly A Farag. "Direct method for shape recovery from polarization and shading." In *Proc. of International Conference on Image Processing*. IEEE, 2012.

[MHM17] Julio Marco, Quercus Hernandez, Adolfo Munoz, Yue Dong, Adrian Jarabo, Min H Kim, Xin Tong, and Diego Gutierrez. "DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging." **36**(6):219, 2017.

[MHN13] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. "Rectifier nonlinearities improve neural network acoustic models." In *Proc. icml*, volume 30, p. 3, 2013.

[MKI04] Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. "Transparent surface modeling from a pair of polarization images." **26**(1):73–82, 2004.

[MSB16] D. Miyazaki, T. Shigetomi, M. Baba, R. Furukawa, S. Hiura, and N. Asada. "Surface normal estimation of black specular objects from multiview polarization images." *Optical Engineering*, **56**(4):041303, 2016.

[MTH03] Daisuke Miyazaki, Robby T Tan, Kenji Hara, and Katsushi Ikeuchi. "Polarization-based inverse rendering from a single view." 2003.

[NNT15] Trung Thanh Ngo, Hajime Nagahara, and R. Taniguchi. "Shape and light directions from shading and polarization." 2015.

[PGC17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in pytorch." 2017.

[PLD18] Jinshan Pan, Yang Liu, Jiangxin Dong, Jiawei Zhang, Jimmy Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang. "Physics-based generative adversarial models for image restoration and beyond." *arXiv preprint arXiv:1808.00605*, 2018.

[Pol17] PolarM polarization camera. "http://www.4dtechnology.com/products/polarimeters/polarcam/." 2017.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *Proc. of International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[SHI18] SHINING 3D scanner. "https://www.einscan.com/einscan-se-sp." 2018.

[SHW18] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. "Deep End-to-End Time-of-Flight Imaging." 2018.

[SMW19] Boxin Shi, Zhipeng Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. "A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo." **41**(2):271–284, 2019.

[SRT16] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. "Linear depth estimation from an uncalibrated, monocular polarisation image." 2016.

[SRT18] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. "Height-from-Polarisation with Unknown Lighting or Albedo." 2018.

[SSO18] Guanya Shi, Xichen Shi, Michael O'Connell, Rose Yu, Kamyar Azizzadenesheli, Animashree Anandkumar, Yisong Yue, and Soon-Jo Chung. "Neural lander: Stable drone landing control using learned dynamics." *arXiv preprint arXiv:1811.08027*, 2018.

[SSS17] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. "Deep photometric stereo network." In *Proc. of International Conference on Computer Vision Workshops*, 2017.

[TM18] Tatsunori Taniai and Takanori Maehara. "Neural inverse rendering for general reflectance photometric stereo." In *Proc. of International Conference on Machine Learning*, 2018.

[TSZ17] Silvia Tozza, William A. P. Smith, Dizhong Zhu, Ravi Ramamoorthi, and Edwin R. Hancock. "Linear Differential Constraints for Photo-polarimetric Height Estimation." 2017.

[Wol97] Lawrence B. Wolff. "Polarization vision: A new sensory approach to image understanding." *Image Vision Computing*, **15**(2):81–93, 1997.

[YTL18] Luwei Yang, Feitong Tan, Ao Li, Zhaopeng Cui, Yasutaka Furukawa, and Ping Tan. "Polarimetric Dense Monocular SLAM." 2018.