Learning Photometric Stereo via Manifold-based Mapping

Anonymous VCIP submission

***Abstract*—Three-dimensional reconstruction technologies are fundamental problems in computer vision. Photometric stereo recovers the surface normals of a 3D object from varying shading cues, prevailing in its capability for generating fine surface normal. In recent years, deep learning-based photometric stereo methods are capable of improving the surface-normal estimation under general non-Lambertian surfaces, due to its powerful fitting ability on the non-Lambertian surface. These state-of-the-art methods however usually regress the surface normal directly from the high-dimensional features, without exploring the embedded structural information. This results in the under-utilization of the information available in the features. Therefore, in this paper, we propose an efficient manifold-based framework for learning-based photometric stereo, which can better map combined high-dimensional feature spaces to low-dimensional manifolds. Extensive experiments show that our method, learning with the low-dimensional manifolds, achieves more accurate surface-normal estimation, outperforming other state-of-the-art methods on the challenging DiLiGenT benchmark dataset.**

1. Introduction

Photometric stereo [1] aims to recover the surface normal of a target from a set of images, captured under different lighting directions. In the past few decades, researchers have worked to apply photometric stereo to a wide range of real-world non-Lambertian objects. Many methods have addressed this problem by applying the outlier rejection methods [2], [3] and modeling sophisticated reflectance methods [4], [5]. Inspired by the success of deep-learning frameworks for various computer vision tasks, researchers have been investigating the learning approaches to photometric stereo [6], [7]. These learning-based methods can better solve the condition of non-Lambertian reflectance through the power of data-driven learning.

However, these deep-learning-based methods [7], [8], [9] fail to explicitly explore the intrinsic data structure of the images' high-dimensional features in the deep neural networks. Usually, these models directly receive a series of feature maps from multiple-input images, which discard a large amount of the features from the input. This greatly reduces the utilization of information and affects the estimation accuracy.

To improve the accuracy of surface-normal estimation, we provide a more efficient solution for learning-based photometric stereo, which can better reconstruct the surface normals from the high-dimensional combined image features from a number of images. In this paper, we propose an end-to-end deep neural framework, as shown in Fig.1. The first stage of the network consists of residual blocks [10], to extract the features from images under different illumination directions. In the second stage, a manifold-based mapping is applied to reconstruct a low-dimensional manifold from the combined extracted features. In our network, isometric feature mapping (ISOMap) [11], a nonlinear dimensionality reduction technique, is employed. In the final stage of our network, based on the low-dimensional manifold, a fully convolutional network [12] is employed to function as a regressor to infer the surface normals of the target. Experiments demonstrate that the proposed network is more efficient and accurate than existing state-of-the-art approaches.

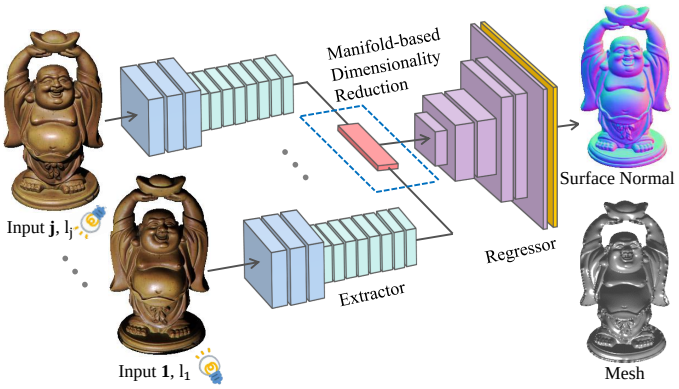

Figure 1. An overview of the proposed network. Given an arbitrary number of images under different light directions, an extractor extracts feature representations from the input images. Then, a manifold method, ISOMap, is employed to map the high-dimensional combined features to a low-dimensional manifold. Finally, a regressor infers the normal as well as the 3D surface of the object concerned.

2. Related Work

Assuming that a pixel in the visual observation of a real-world object m_j can be modeled by general bidirectional reflectance distribution functions (BRDFs) ρ, which is associated with the surface normal $\boldsymbol{n} \in \mathbb{R}^3$, under illumination direction $\boldsymbol{l_j} \in \mathbb{R}^3$. The imaging model can be expressed as follows:

$$m_j = e_j \rho\left(\boldsymbol{n}, l_j\right) \max\left(\boldsymbol{n}^\top \boldsymbol{l}_j, 0\right) + \epsilon_j \tag{1}$$

where $\max\left(\boldsymbol{n}^\top \boldsymbol{l}_j, 0\right)$ accounts for the attached shadows, and ϵ represents the noise and global illumination effect like inter-reflections and image noise.

The goal of photometric stereo [1] is to recover the surface orientation at every pixel position from a combination of reflectance and illuminations in multiple images. To extend the photometric stereo to work with non-Lambertian surfaces in practical use, researchers have investigated two kinds of strategies, namely outlier rejection methods, sophisticated reflectance models, following the reference [13].

Recently, some deep-learning-based methods have been introduced to the field of surface normal reconstruction [14], [15]. Santo *et al.* first investigated the deep photometric stereo network (DPSN) [6], by using fully connected layers to regress per-pixel surface normals, based on a fixed number of images (96 in the method), with different illumination directions. However, DPSN directly reduces the high-dimensional feature of 96 input images from the dimensionality 2048 to a 3-dimensional output with a single fully-connected layer, without exploring the embedded information in the high-dimensional data. Chen *et al.* [8] proposed a method, called PS-FCN, which uses a fully convolutional network to regress the surface normals. Furthermore, Chen *et al.* [7] proposed an SDPS-Net to determine both the surface normals and light directions of an object, for uncalibrated photometric stereo. These two methods can take an arbitrary number of images as their input, by applying channel max pooling. However, because of the use of max pooling, most of the features are not considered in the process of nonlinear dimensionality reduction. In this paper, we apply an efficient manifold-based mapping method. ISOMap [11] to reconstruct the low-dimensional manifold, considering the embedded structural information and improving the utilization of the information available in the input images.

3. Proposed Method

3.1. Network architecture

As in the architecture shown in Fig.1, the first 118 layers of the ResNet-152 model [10] function as an extractor network, denoted as $f_{\text{extractor}}$. Note that the light direction $\boldsymbol{l}$ is expanded to the same spatial size as the input image and is concatenated with the input image patch. We obtain a combined high-dimensional feature of dimensionality at the end of the extractor network $\boldsymbol{\Psi} \in \mathbb{R}^{jd \times h' \times w'}$, where d is 256 in our network, which is the number of feature maps generated, and $h' = \frac{1}{8}h$ and $w' = \frac{1}{8}w$, as follows:

$$\boldsymbol{\Psi} = f_{\text{extractor}}(\boldsymbol{I}; \theta_{\text{extractor}}), \tag{2}$$

where $\theta_{\text{extractor}}$ are the learnable parameters in $f_{\text{extractor}}$.

We then apply the ISOMap manifold-learning method to map the combined feature $\boldsymbol{\Psi} \in \mathbb{R}^{jd \times h' \times w'}$ to a low-dimensional manifold $\boldsymbol{\Phi} \in \mathbb{R}^{d \times h' \times w'}$, with the details discussed in Section 3.2.

Finally, the regressor network takes the manifold output $\boldsymbol{\Phi}$ as its input, and estimates a normal map of the target object in the images. We design a 6-layer fully convolutional network $f_{\text{regressor}}$, with an L2-normalization layer f_{norm} (highlighted in yellow in Fig.1), at the end of the network, to regress the surface normal at the i-th pixel position $\boldsymbol{n_i} \in \mathbb{R}^{c' \times h \times w}$,where c' is 3 to represent the x, y, and z directions of the surface normal, as follows:

$$\boldsymbol{n_i} = f_{\text{norm}}(f_{\text{regressor}}(\boldsymbol{\Phi}; \theta_{\text{regressor}})), \tag{3}$$

where $\theta_{\text{regressor}}$ are the learnable parameters in $f_{\text{regressor}}$.

The loss function employed in our network is the commonly used cosine similarity loss, defined as follows:

$$\mathcal{L}_{\text{normal}} = \frac{1}{hw} \sum_{i}^{hw} \left(1 - \boldsymbol{n_i}^{\top} \tilde{\boldsymbol{n}}_{\boldsymbol{i}}\right) \tag{4}$$

where hw is the number of pixels in each image, and $\boldsymbol{n_i}$ and $\tilde{\boldsymbol{n}}_{\boldsymbol{i}}$ represent the estimated normal and the ground-truth normal, respectively, at the i-th pixel.

3.2. Manifold-based dimensionality reduction and training strategy

Previous methods [7], [8] applied the channel max pooling to generate low-dimensional features for regression. Compared with a convolutional layer, it allows to having an arbitrary number of images as input. However, the method leads to the loss of most of the feature information, and only retains the maximum response results. Consequently, accuracy will be reduced.

In this paper, we focus on devising a more efficient learning-based feature fusion approach. We employ the manifold-learning method, ISOMap [11], to map high-dimensional features to a low-dimensional manifold, where the mapping fully takes advantage of the embedded data structure. Different from channel max pooling, which only retains the maximum response feature values, ISOMap can generate an effective low-dimensional embedding of a set of high-dimensional data points. Its effectiveness has been shown in learning-based regression tasks [16].

Existing manifold-learning methods, such as ISOMap [11], Principal Component Analysis (PCA) [17], truncate the backpropagation process of tensors, which will cause the extractor to be untrained. To avoid this, we apply a two-step training strategy. We first pre-train the extractor network, as in [8], but give up the weights of the regressor. Then, we retrain the regressor with ISOMap, using the extractor with the fixed pretrained weights. The number of input images stays the same in the pretraining and retraining processes.

4. Experiments

4.1. Dataset

Following [8], we employed the MERL dataset [19] to render the publicly available synthetic Blobby and Sculpture datasets for training [20]. We randomly split the samples in the dataset into a ratio of approximately 99 : 1, for training and validation. In testing, we applied the DiLiGenT dataset [13], which is the most commonly used real-world dataset

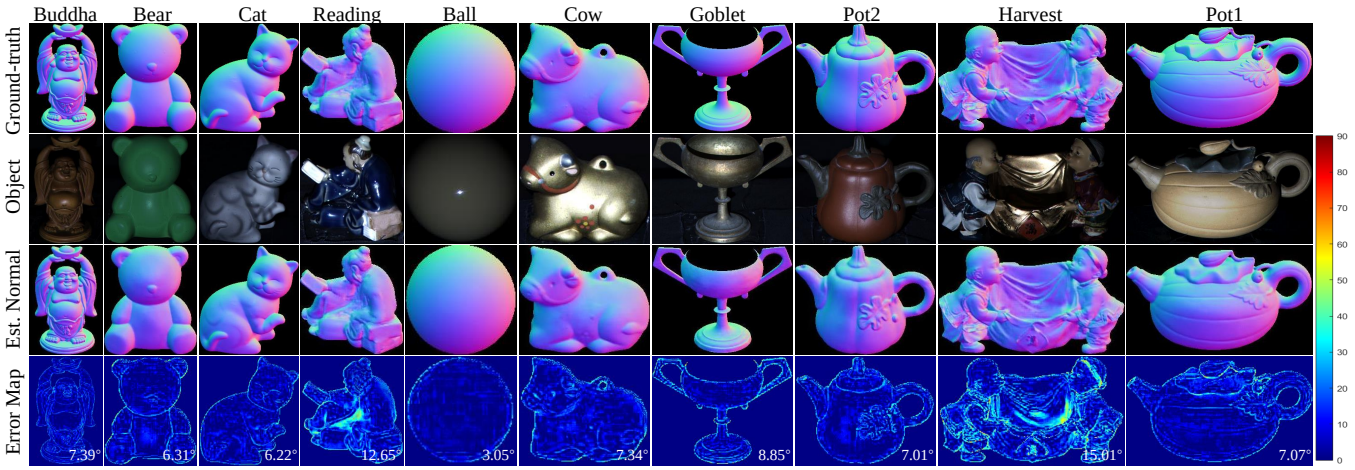

Figure 2. Visual results of our method, with the contrast of the observation images adjusted for easy viewing. The numbers in the error maps represent the MAE in degrees.

for evaluating photometric stereo algorithms. The dataset consists of 10 different objects with different scales of non-Lambertian reflectance.

4.2. Network analysis

We analyze our proposed deep neural network using the validation set in this section. We first compare the performance of our network with the use of different dimensionality reduction methods, including max pooling (Max-p) [8], 1 × 1 Conv, ISOMap [11], and PCA [17]. It is worth noting that the 1 × 1 Conv needs the input dimension fixed. This means that the results exist only when the number input images is the same for training and testing. We also evaluate the influence of the number of input images, used in training and testing, on the performance. The average performances of our proposed method are reported in Table 1. The performances are measured in terms of the mean angular error (MAE), which is calculated as the mean of $cos^{-1}(\tilde{\boldsymbol{n}}_{\boldsymbol{i}} \cdot \boldsymbol{n}_{\boldsymbol{i}})$, in degrees.

We first discuss the effectiveness of the different dimensionality reduction methods. It can be seen that ISOMap achieves a better performance than PCA, because PCA is a simply linear subspace method, while ISOMap is a nonlinear method. For max pooling, its performance is inferior to other methods, except when the number of input images used for traning and testing is 16 and 8, respectively. This method does not explicitly explore the data structure and discards most of the features, but retains only the maximum response. We also note that the results based on 1× 1 Conv are sub-optimal. These experiment results illustrate the effectiveness of 1 × 1 Conv in dimensionality reduction, because this method can handle the feature and space structures simultaneously. However, the number of input images used for training and testing must be the same.

We also compare the performance of the different dimensionality reduction methods, when different numbers of input images are used. In general, all the methods achieve better performance, when more images are used for training, *i.e.*, using 32 images outperforms the use of 16 images, except when only 8 input images are used for testing. This is because the patterns learned by using 16 input images may be closer than using 32 input images in training, to the patterns generated using 8 input images during testing. Moreover, when compared with max pooling, we can see that our method has its average MAE decreased at a higher rate, when the number of input images increases.

TABLE 1. PERFORMANCES OF OUR METHODS USING DIFFERENT DIMENSIONALITY REDUCTION METHODS, WITH DIFFERENT NUMBERS OF INPUTS USED IN TRAINING AND TESTING..

Methods	The number of input images used					
	Training	Testing				
		8	16	32	48	64
Max-p [8]	16	**11.01**	10.03	9.67	9.56	9.49
	32	11.17	9.92	6.51	9.42	9.35
1×1 Conv	16	×	**9.88**	×	×	×
	32	×	×	9.60	×	×
ISOMap [11]	16	11.20	10.11	9.53	9.34	9.05
	32	11.04	9.98	**9.44**	**9.22**	**8.97**
PCA [17]	16	11.80	10.36	9.78	9.46	9.27
	32	11.54	10.30	9.65	9.38	9.21

4.3. Benchmark comparisons

We compare our proposed network against the recently proposed state-of-the-art learning-based methods, including PS-FCN [8] and DPSN [6], as well as non-learning-based methods. Quantitative and qualitative results on the DiLiGenT benchmark [13] are shown in Table 2, and Fig.2, respectively.

We can see that our network outperforms those existing learning-based method, with an average MAE of 8.09°. For those simple objects, like "Ball" and "Cat" with Lambertian surface, the non-learning-based method [4] achieves the best performance. However, our method achieves particularly

TABLE 2. COMPARISON OF DIFFERENT METHODS, IN TERMS OF THE MAE IN DEGREES, ON THE DILIGENT DATASET. ALL THE METHODS ARE EVALUATED WITH 96 IMAGES (TRAINED WITH 32 IMAGES IN OUR NETWORK).

Method	ball	cat	pot1	bear	pot2	buddha	goblet	reading	cow	harvest	Avg.
L2 [1]	4.10	8.41	8.89	8.39	14.65	14.92	18.50	19.80	25.60	30.62	15.39
WG10 [3]	2.06	6.73	7.18	6.50	13.12	10.91	15.70	15.39	25.89	30.01	13.35
HM10 [18]	3.55	8.40	10.85	11.48	16.37	13.05	14.89	16.82	14.95	21.79	13.22
AZ08 [5]	2.71	6.53	7.23	**5.96**	11.03	12.54	13.93	14.17	21.48	30.50	12.61
ST14 [4]	**1.74**	**6.12**	**6.51**	6.12	8.78	10.60	10.09	13.63	13.93	25.44	10.30
DPSN [6]	2.02	6.54	7.05	6.31	7.86	12.68	11.28	15.51	8.01	16.86	9.41
PS-FCN [8]	2.82	6.16	7.13	7.55	7.25	7.91	**8.60**	13.33	**7.33**	15.85	8.39
Ours	3.05	6.22	7.07	6.31	**7.01**	**7.39**	8.85	**12.65**	7.34	**15.01**	**8.09**

good performance on objects with complicated structures and strongly non-Lambertian surfaces, such as "Reading" and "Harvest".

5. Conclusions

This paper proposes a deep learning method for photometric stereo. The proposed method can take an arbitrary number of images as input and regresses a fine surface normal. We apply the ISOMap method for mapping the combined high-dimensional features from the extractor network to a low-dimensional manifold, rather than using the max pooling fusion method. Experiments on a public benchmark show that our method outperforms existing approaches on objects with complicated structures and strongly non-Lambertian surfaces. We have further evaluated our method on the utilization rate, by measuring how well our method can extract useful information from input images, when the number is increasing. Based on the experiment results, we can conclude that our method can achieve state-of-the-art performance, when the number of input images increases to a certain level.

References

[1] R. J Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 139–144, 1980.

[2] Daisuke Miyazaki, Kenji Hara, and Katsushi Ikeuchi, "Median photometric stereo as applied to the segonko tumulus and museum objects," *International Journal of Computer Vision*, vol. 86, no. 2-3, pp. 229, 2010.

[3] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Asian Conference on Computer Vision*, 2010, pp. 703–717.

[4] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1078–1091, 2014.

[5] Neil Alldrin, Todd Zickler, and David Kriegman, "Photometric stereo with non-parametric and spatially-varying reflectance," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[6] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita, "Deep photometric stereo network," in *IEEE International Conference on Computer Vision Workshop*, 2017, pp. 501–509.

[7] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong, "Self-calibrating deep photometric stereo networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8739–8747.

[8] Guanying Chen, Kai Han, and Kwan-Yee K Wong, "Ps-fcn: A flexible learning framework for photometric stereo," in *European Conference on Computer Vision*. Springer, Cham, 2018, pp. 3–19.

[9] Yakun Ju, Kim-man Lam, Yang Chen, Lin Qi, and Junyu Dong, "Pay attention to devils: A photometric stereo network for better details," in *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI*, 2020.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Joshua B Tenenbaum, Vin De Silva, and John C Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[13] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2018.

[14] Tatsunori Taniai and Takanori Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *International Conference on Machine Learning*, 2018, pp. 4864–4873.

[15] Yakun Ju, Xinghui Dong, Yingyu Wang, Lin Qi, and Junyu Dong, "A dual-cue network for multispectral photometric stereo," *Pattern Recognition*, vol. 100, pp. 107162, 2020.

[16] Yan Jia, Yinqiang Zheng, Lin Gu, Art Subpa-Asa, Antony Lam, Yoichi Sato, and Imari Sato, "From rgb to spectrum for natural scenes via manifold-based mapping," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4705–4713.

[17] Svante Wold, Kim Esbensen, and Paul Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[18] Tomoaki Higo, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Consensus photometric stereo," in *2010 IEEE computer society conference on computer vision and pattern recognition*, 2010, pp. 1157–1164.

[19] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan, "A data-driven reflectance model," *ACM Transactions on Graphics*, 2003.

[20] Micah K Johnson and Edward H Adelson, "Shape estimation in natural illumination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2011, pp. 2553–2560.