Deep Discrete Wavelet Transform Network for Photometric Stereo

Yakun Ju*, Muwei Jian†, Cong Zhang*, Yeqi Hu‡, Kin-Man Lam*
*Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR
Email: kelvin.yakun.ju@gmail.com, cong-clarence.zhang@connect.polyu.hk, enkmlam@polyu.edu.hk
†School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China
Email: jianmuweihk@163.com
‡School of Computer Science and Technology, Ocean University of China, Qingdao, China
Email: huyeqi@stu.ouc.edu.cn

***Abstract*—Photometric stereo aims to estimate the per-pixel surface normal map of 3D objects via changing the illuminated light directions. Prevalent methods adopt deep neural networks to extract the shading cue features and reconstruct the surface normals. However, previous methods do not consider the frequency of the surface structure, *i.e.*, the complexity of the shape. Simply applying a trained network to all kinds of objects often leads to inter-frequency conflicts and blur in surface normal estimation. This paper presents a discrete wavelet transform-based photometric stereo network (DWTPS-Net) to handle the input photometric stereo images in both the spatial and frequency domains. In DWTPS-Net, we extract shading features from images and also decompose the images using discrete wavelet transform (DWT), which can preserve spatial information naturally, to better extract high-frequency information. We design separate CNN-based feature-extraction modules for the input images and for the different frequency information of the input images via DWT. Ablation studies and experiments on a widely used benchmark dataset show that DWTPS-Net achieves superior performance in surface normal estimation, in terms of mean angular error metric.**

I. Introduction

Three-dimensional (3D) shape recovery is a vital problem in computer vision since it will improve the understanding of two-dimensional (2D) images. Photometric stereo [1], [2] is a 3D shape recovery method that measures the pixel-wise 3D surface normal $\boldsymbol{n}$ of the intensity o of an object under changing light direction $\boldsymbol{l}$ from the view $\boldsymbol{v}$, as follows:

$$\boldsymbol{o} = \boldsymbol{\rho}\left(\boldsymbol{n}, \boldsymbol{l}, \boldsymbol{v}\right) \max\left(\boldsymbol{n}^{\top}\boldsymbol{l}, 0\right), \tag{1}$$

where $\max\left(\boldsymbol{n}^{\top}\boldsymbol{l}, 0\right)$ reveals the attached shadows. Traditional photometric stereo methods [3], [4], [5] attempt to solve the nonlinear relationship between image shading cues and surface normal orientations affected by non-Lambertian reflectance $\boldsymbol{\rho}$. However, these traditional methods are accurate for a limited category of materials and suffer from unstable optimization.

Recently, photometric stereo methods based on deep learning have been proposed [6], [7], [8] and demonstrated robust and superior performance in recovering surface normals of non-Lambertian objects [9], [10]. However, previous learning-based methods ignore the ability of networks to handle surface structures of different frequencies. For regions with complex-structure, such as crinkles and edges on objects, patch-based methods always fail to predict and lead to blurring [11], [12], [13]. There are two reasons for this. First, they process the input images in a patch-wise manner, which barely handles steeply changed pixel' values, *i.e.*, high-frequency surface structures [11] and spatially varying materials [14]. Second, the widely used L2/cosine losses in reconstruction focus more on recovering low-frequency global structures [15]. Although previous work [11] pays more attention to high-frequency surface regions, it is limited in spatially varying materials and hard to optimize.

979-8-3503-3959-8/23/$31.00 © 2023 IEEE

To address the aforementioned issues, we propose a discrete wavelet transform-based photometric stereo network (DWTPS-Net) to better simultaneously tackle high-frequency structural details and low-frequency global information. DWTPS-Net employs wavelet transform to extract the shading cue features in the frequency domain and spatial domain. As shown in Fig. 1, we convert the normalized input images into the wavelet domain using 2D discrete wavelet transform (DWT) [16], [17], where the images are decomposed into a low-frequency band (LF), a vertical high-frequency band (VH), a horizontal high-frequency band (HH), and a diagonal high-frequency band (DH). Then, we concatenate VH, HH, and DH as the high-frequency information input (HF). For the different frequency information of the images (original, low-frequency, and high-frequency), we design separate convolutional layers to extract the features, which are then fused in the regression stage.

To the best of our knowledge, we propose the first deep photometric stereo network that considers both the spatial domain and frequency domain of the input via DWT. The ablation study demonstrates the effectiveness of the spatial-frequency domains extraction module, and benchmark comparisons show the superior performance of our method.

II. Proposed Method

In this Section, we will present our proposed method. Before we introduce wavelet decomposition and network architecture, we first explain some baseline operations for deep learning-based photometric stereo methods.

A. Baseline operations

1) Observation normalization: A CNN-based photometric stereo framework may fail to estimate an input with spatially

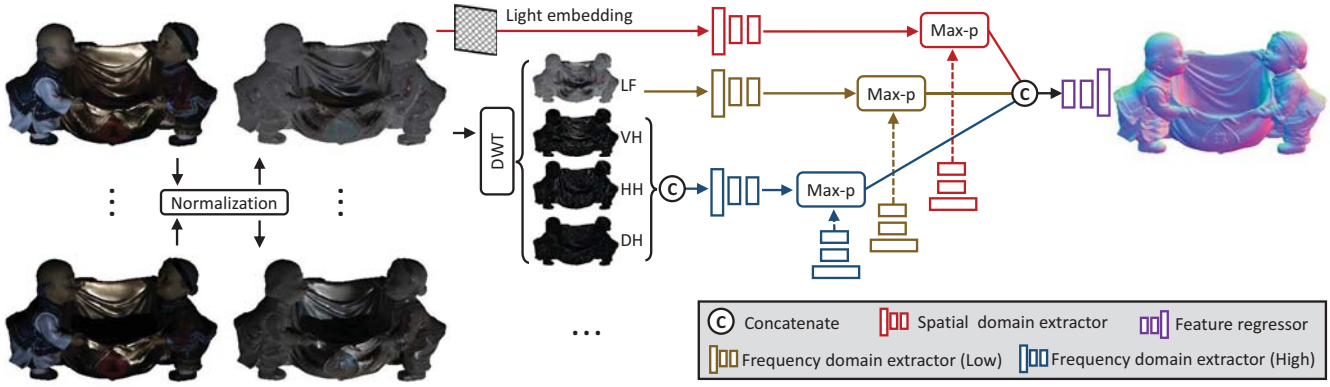

Fig. 1. The framework of our DWTPS-Net. First, we perform the observation normalization operation on the inputs images (Section II-A1). For the normalized images, we embed the light direction into them (Section II-A2). Then, we decompose the normalized images with discrete wavelet transform, forming low-frequency and high-frequency parts in the frequency domain (Section II-B1). The spatial and frequency domains of the input are then extracted by different extractors, followed by max-pooling fusion and a feature regressor (Section II-B2).

varying colors, because it handles the patch-level inputs and a patch with different colors may cause mutual influence. Therefore, we first employ an observation normalization method [14] to remove the impact of spatially varying surface materials. The operation is to normalize each observation by all n observations, as follows:

$$o_i' = \frac{o_i}{\sqrt{o_1^2 + o_2^2 + \cdots + o_n^2}}, \ i \in \{1, 2, \cdots, n\}, \qquad (2)$$

where o_i and o_i' represent a pixel value in the i_{th} original observation and the normalized observation, respectively. Under the assumption of Lambertian reflectance, the reflectance $\boldsymbol{\rho}$ in Eq. 1 can be removed. With this baseline operation, subsequent DWT can avoid extracting false high-frequency information from the sharply changing changed surface colors.

2) Light direction embedding: As a calibrated photometric stereo method, the light direction of each input image should be known and input into the network. However, an incident light direction is a xyz 3-dimensional vector $\boldsymbol{l} \in \mathbb{R}^3$, which cannot be fused with the input images. Therefore, we repeat each light direction $\boldsymbol{l_i}$ to form a 3-channel image having the same spatial dimension as the input image $\boldsymbol{O_i} \in \mathbb{R}^{H \times W \times 3}$, and can be concatenated with input images, forming the combined feature $\boldsymbol{\Phi} \in \mathbb{R}^{H \times W \times 6}$.

3) Arbitrary features fusion: A unique problem faced by deep photometric stereo methods is the unfixed number and order-agnostic input images. The CNN-based structure can hardly handle a variable number of inputs, since it needs the input to keep a fixed number of channels during training and testing [7]. Therefore, we apply a max-pooling operation, which retains the strongest features (most activated value) from all the input features, to aggregate any number of extracted features into a single feature that can be optimized by backpropagation.

B. DWTPS-Net

1) Wavelet decomposition: As shown in Fig. 1, our main contribution is to decompose the normalized images into different frequency bands via DWT and extract the features in both the spatial domain and frequency domain (both low and high).

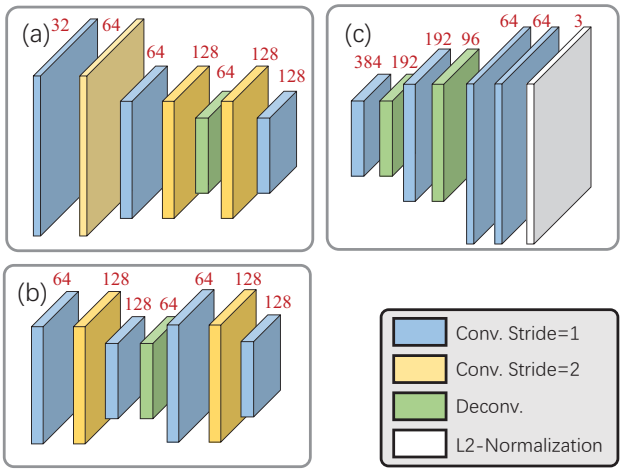

Fig. 2. Network architecture of (a) the spatial domain extractor, (b) the frequency domain extractor, and (c) the feature regressor. The red numbers represent the size of channels after the convolutional operation.

We adopt the 2D discrete wavelet transform with the Haar wavelet filters as the basis to decompose images into multiple wavelet sub-bands with different frequency components. The transform produces four sub-bands, including low-frequency band (LF), vertical high-frequency band (VH), horizontal high-frequency band (HH), and diagonal high-frequency band (DH). VH, HH, and DH are the high-frequency content along the vertical, horizontal, and diagonal directions, respectively. Therefore, we concatenate them into a single feature representing high-frequency information (HF). It is worth noting that LF, VH, HH, and DH have half the resolution of the input normalized image.

2) Network architecture: Our DWTPS-Net is a multi-branch Siamese network consisting of three components, the shared-weight extractors (for both spatial and frequency domain), the max-pooling feature aggregator (see Section II-A3), and the normal regressor. To largely exclude the influence of other factors and verify the effectiveness of the main modules in our proposed network, all the sub-networks follow the

simple fully convolutional layers, as described in Fig. 2. Due to the different resolutions of the spatial-domain input and the frequency-domain input, we design two extractors, for extracting features from $\mathbb{R}^{H\times W}$ and $\mathbb{R}^{H/2\times W/2}$ to $\mathbb{R}^{H/4\times W/4}$, respectively. All the convolution layers employ 3×3 kernel size. An L2-normalization layer is appended at the end of the regressor to produce a unit normal map. The max-pooling operations are used to possess an arbitrary number of features, between the extractor and regressor.

C. Learning procedures

We optimize the proposed SWTPS-Net by minimizing the cosine loss function $\mathcal{L}$, as follows:

$$\mathcal{L} = \frac{1}{P}\sum_{p}^{P}(1 - \tilde{\boldsymbol{n}}_{\boldsymbol{p}} \cdot \boldsymbol{n}_{\boldsymbol{p}}), \tag{3}$$

where p represents the index of the pixel location on the image and P is the total number of pixels. If the estimated surface normal $\tilde{\boldsymbol{n}}_{\boldsymbol{p}}$ has a similar orientation to the ground-truth $\boldsymbol{n}_{\boldsymbol{p}}$ at pixel p, $\tilde{\boldsymbol{n}}_{\boldsymbol{p}} \cdot \boldsymbol{N}_{\boldsymbol{p}}$ will be close to one and Eq. (3) will approach zero.

DWTPS-Net is optimized using the default Adam optimizer (β_1=0.9, β_2=0.999) and trained on the widely used Blobby and Sculpture shape dataset rendered with MERL BRDF [7], [14], [11], [8], [18], [19]. We train the model using a batch size of 32 for 30 epochs. The number of input images for training is 32, with the size 32×32, and can be any number during testing.

III. Experiments

In this section, we present the experimental results and analysis of the proposed DWTPS-Net. We use the mean angular error (MAE) in degrees to quantitatively evaluate the performance of our method and other state-of-the-art methods, as follows:

$$\mathrm{MAE} = \frac{1}{P}\sum_{p}^{P}\cos^{-1}\left(\tilde{\boldsymbol{n}}_{\boldsymbol{p}} \cdot \boldsymbol{n}_{\boldsymbol{p}}\right). \tag{4}$$

A. Ablation studies

To test the effect of the proposed frequency-domain decomposition, we first quantitatively compare our proposed network with or without DWT and different settings of the DWT outputs. We also test the effect of the normalization operation. As tabulated in Table I, we conducted these ablation experiments on the widely used DiLiGenT benchmark (96 input images) [20].

As tabulated in Table I, (a) is our DWTPS-Net, (b) removes all the frequency-domain inputs (also without the frequency-domain extractors). (c) and (d) discard the LF input and HF input, respectively. While (e) cancels the concatenation operation for fusing VH, HH, and DH. Therefore, there are four frequency-domain extractors in the full model. We further evaluate the condition without observation normalization in (f), as well as repeating (b), (c), and (d) without observation normalization, denoted as IDs (g), (h), and (i), respectively.

Discussion 1: Compared with (b), we can see that the proposed frequency-domain feature extraction improves the estimation accuracy of surface normals. This is because the network can better learn the complex structural features in crinkle and edge regions. We further test the effect of the low-frequency part and the high-frequency part in (c) and (d). When the network discards the low-frequency information, the performance only decreases slightly. However, the performance becomes worse, if the high-frequency information is discarded. The results show that the combined high-frequency information is much more helpful in recovering the surface normals. This can be explained by the fact that the spatial inputs (normalized images) contain more low-frequency global information but lacks high-frequency details. Moreover, (e) shows slight improvement when we separately extract features in the three high-frequency bands along different directions. This is simply because it uses more parameters to learn the features, thus greatly increasing the computational burden.

TABLE I. Results of the ablation studies of the proposed method, based on the average MAE of ten objects from the DiLiGenT benchmark [20].

ID	Methods	MAE
(a)	Proposed method	7.03
(b)	w/o DWT	7.66
(c)	w/o LF	7.09
(d)	w/o HF	7.56
(e)	Separate bands (VH, HH, DH)	**7.01**
(f)	w/o Normalization	8.32
(g)	w/o (Normalization + DWT)	8.20
(h)	w/o (Normalization + LF)	8.36
(i)	w/o (Normalization + HF)	8.19

Discussion 2: In addition, we explore the above-mentioned variants under the condition without observation normalization. The results are surprising. Comparing (f) to (g), we can find that the proposed frequency-domain extractors suffer from performance drop when photometric stereo images are not normalized (8.32 *v.s.* 8.20). This suggests that the decomposed high-frequency content is sensitively affected by spatially varying surface materials, where the sharply changing colors cause drastic differences in the intensity of photoed pixels. To further demonstrate this, we add (h) and (i) at the end. It can be seen that retaining high-frequency information only makes the performance worse, while retaining low-frequency information can slightly improve the results.

B. Benchmark comparison

The DiLiGenT dataset [20] is a widely used benchmark for evaluating photometric stereo methods. It contains 10 objects of different shapes and complex non-Lambertian surfaces, illuminated under 96 light directions. Therefore, we compare our DWTPS-Net with traditional algorithms and state-of-the-art deep learning-based methods on DiLiGenT. The results are summarized in Table II. We also show some visualization examples in Fig. 3.

As tabulated in Table II, our DWTPS-Net achieves state-of-the-art performance on the DiLiGenT benchmark. It is worth noting that we only compare the methods trained on the MERL BRDF dataset [24] for fairness. We can see that our method particularly works well for complex objects, such as Buddha, Harvest, and Reading. As shown in Fig. 3, our method achieves advanced surface-normal estimation for complex objects, such as the Harvest's cloth, the lid of Pot1, and the floral pattern of

TABLE II. COMPARISON OF DIFFERENT METHODS ON THE DILIGENT BENCHMARK [20]. ALL METHODS ARE EVALUATED WITH 96 IMAGES.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Average
LS [1]	4.10	8.39	14.92	8.41	25.60	18.50	30.62	8.89	14.65	19.80	15.39
WG10 [3]	2.06	6.50	10.91	6.73	25.89	15.70	30.01	7.18	13.12	15.39	13.35
HM10 [4]	3.55	11.48	13.05	8.40	14.95	14.89	21.79	10.85	16.37	16.82	13.22
ST14 [5]	**1.74**	6.12	10.60	6.12	13.93	10.09	25.44	6.51	8.78	13.63	10.30
DPSN [6]	2.02	6.31	12.68	6.54	8.01	11.28	16.86	7.05	7.86	15.51	9.41
LMPS [19]	2.40	5.23	9.89	6.11	7.98	8.61	16.18	6.54	7.48	13.68	8.41
PS-FCN [7]	2.82	7.55	7.91	6.16	7.33	8.60	15.85	7.13	7.25	13.33	8.39
Attention-PSN[11]	2.93	4.86	7.75	6.14	6.86	8.42	15.44	6.92	6.97	12.90	7.92
DR-PSN [21]	2.27	5.46	7.84	5.42	7.01	8.49	15.40	7.08	7.21	12.74	7.90
CHR-PSN [22]	2.26	6.35	7.15	5.97	6.05	8.32	15.32	7.04	6.76	12.52	7.77
GPS-Net [12]	2.92	5.07	7.77	5.42	6.14	9.00	15.14	6.04	7.01	13.58	7.81
MT-PS-CNN [23]	2.29	5.79	6.85	5.87	7.48	7.88	13.71	6.92	6.89	11.94	7.56
PS-FCN(Norm.) [14]	2.67	7.72	7.53	4.76	6.72	7.84	12.39	6.17	7.15	10.92	7.39
MF-PSN [18]	2.07	5.83	6.88	5.00	5.90	7.46	13.38	7.20	6.81	12.20	7.27
Ours	2.61	5.97	6.81	4.72	6.71	7.99	12.47	5.96	6.54	10.58	7.03

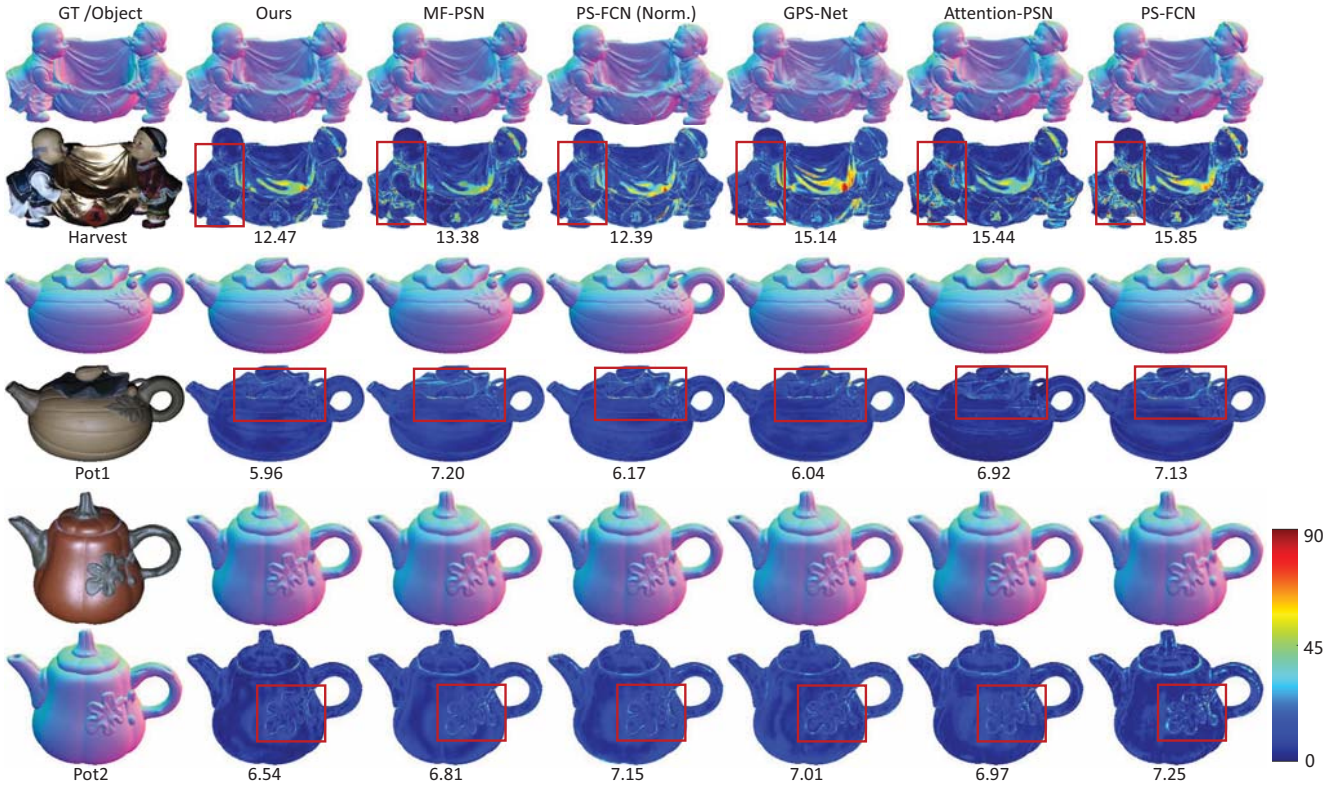

Fig. 3. Visualization results of complex objects from the DiLiGenT benchmark [20]. Red boxes are regions with high-frequency complex structures. Our method significantly outperforms other methods.

Pot2. These results illustrate the effect of the introduced DWT decomposition and frequency-domain extraction.

IV. CONCLUSION

In this paper, we have proposed a discrete wavelet transform-based photometric stereo network. To realize accurate surface-normal reconstruction with details, we introduce wavelet decomposition of photometric stereo images. With the decomposed low-frequency and high-frequency domain information, we further propose different extractors to extract global structures and detailed features from the spatial and frequency domain of input images. Ablation studies demonstrate the effectiveness of the proposed method. Comparisons on the widely used DiLiGenT benchmark show the better performance of our method.

ACKNOWLEDGMENT

The work was supported by the Project of Strategic Importance Fund from The Hong Kong Polytechnic University (No. ZE1X) Key Development Program for Basic Research of Shandong Province (ZR2020ZD44).

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] M. Jian, J. Dong, M. Gong, H. Yu, L. Nie, Y. Yin, and K.-M. Lam, "Learning the traditional art of chinese calligraphy via three-dimensional reconstruction and assessment," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 970–979, 2019.

[3] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Proceedings of the Asian Conference on Computer Vision*. Springer, 2010, pp. 703–717.

[4] T. Higo, Y. Matsushita, and K. Ikeuchi, "Consensus photometric stereo," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1157–1164.

[5] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 1078–1091, 2014.

[6] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep photometric stereo network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 501–509.

[7] G. Chen, K. Han, and K.-Y. K. Wong, "Ps-fcn: A flexible learning framework for photometric stereo," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–18.

[8] Y. Ju, M. Jian, S. Guo, Y. Wang, H. Zhou, and J. Dong, "Incorporating lambertian priors into surface normals measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[9] Q. Zheng, B. Shi, and G. Pan, "Summary study of data-driven photometric stereo methods," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 3, pp. 213–221, 2020.

[10] Y. Ju, K.-M. Lam, W. Xie, H. Zhou, J. Dong, and B. Shi, "Deep learning methods for calibrated photometric stereo and beyond: A survey," *arXiv preprint arXiv:2212.08414*, 2022.

[11] Y. Ju, K. Lam, Y. Chen, L. Qi, and J. Dong, "Pay attention to devils: A photometric stereo network for better details," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2020, pp. 694–700.

[12] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi, "Gps-net: Graph-based photometric stereo network," in *Proceedings of Advances in Neural Information Processing Systems*, 2020, p. 33.

[13] Y. Ju, B. Shi, M. Jian, L. Qi, J. Dong, and K.-M. Lam, "Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention," *International Journal of Computer Vision*, vol. 130, no. 12, pp. 3014–3034, 2022.

[14] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Deep photometric stereo for non-lambertian surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 129–142, 2020.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[16] M. J. Shensa *et al.*, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on signal processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[17] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao, "Wavefill: A wavelet-based generation network for image inpainting," in *Proceedings of the International Conference on Computer Vision*, 2021, pp. 14 114–14 123.

[18] Y. Liu, Y. Ju, M. Jian, F. Gao, Y. Rao, Y. Hu, and J. Dong, "A deep-shallow and global–local multi-feature fusion network for photometric stereo," *Image and Vision Computing*, vol. 118, p. 104368, 2022.

[19] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita, "Learning to minify photometric stereo," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7568–7576.

[20] B. Shi, Z. Mo, Z. Wu, D. Duan, S.-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 271–284, 2019.

[21] Y. Ju, J. Dong, and S. Chen, "Recovering surface normal and arbitrary images: A dual regression network for photometric stereo," *IEEE Transactions on Image Processing*, vol. 30, pp. 3676–3690, 2021.

[22] Y. Ju, Y. Peng, M. Jian, F. Gao, and J. Dong, "Learning conditional photometric stereo with high-resolution features," *Computational Visual Media*, vol. 8, pp. 105–118, 2022.

[23] Y. Cao, B. Ding, Z. He, J. Yang, J. Chen, Y. Cao, and X. Li, "Learning inter-and intraframe representations for non-lambertian photometric stereo," *Optics and Lasers in Engineering*, vol. 150, p. 106838, 2022.

[24] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 759–769, 2003.