Pay Attention to Devils: A Photometric Stereo Network for Better Details

Yakun Ju[1] , **Kin-Man Lam**[2] , **Yang Chen**[1] , **Lin Qi**[1] and **Junyu Dong**[1*]
[1]Department of Computer Science and Technology, Ocean University of China
[2]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University
{juyakun, chenyang8484}@stu.ouc.edu.cn, kin.man.lam@polyu.edu.hk, {qilin, dongjunyu}@ouc.edu.cn

Abstract

We present an attention-weighted loss in a photometric stereo neural network to improve 3D surface recovery accuracy in complex-structured areas, such as edges and crinkles, where existing learning-based methods often failed. Instead of using a uniform penalty for all pixels, our method employs the attention-weighted loss learned in a self-supervise manner for each pixel, avoiding blurry reconstruction result in such difficult regions. The network first estimates a surface normal map and an adaptive attention map, and then the latter is used to calculate a pixel-wise attention-weighted loss that focuses on complex regions. In these regions, the attention-weighted loss applies higher weights of the detail-preserving gradient loss to produce clear surface reconstructions. Experiments on real datasets show that our approach significantly outperforms traditional photometric stereo algorithms and state-of-the-art learning-based methods.

1 Introduction

3D shape recovery is a fundamental problem in computer vision [Jian *et al.*, 2019]. It is well known that photometric methods prevail in recovering fine details of objects and play an essential role in machine vision and highly accurate 3D recovery. Photometric stereo recovers the dense surface normal of the object under different illumination directions, with a fixed camera [Woodham, 1980]. The multiple images, under different illumination directions, provide the varying shading cues to recover 3D surface normals based on Lambertian assumption. To make photometric stereo applicable to real-world objects, subsequent methods focus more on non-Lambertian surfaces with more flexible reflectance functions [Chung and Jia, 2008; Ruiters and Klein, 2009].

Recently, deep learning frameworks have shown great success in handling non-Lambertian surfaces, because of their powerful fitting ability [Santo *et al.*, 2017; Chen *et al.*, 2018], which achieves state-of-the-art performance. However, as shown in Figure 1, the majority of the errors still exist in some complex-structured regions, such as the edges and crinkles of the objects. Unfortunately, these areas are where we focus on and require to be reconstructed accurately. The failure in predicting these areas is caused by the estimated surface normal, which is not complex enough (blurry). This result is due to the use of unsuitable receptive fields in the convolutional networks and the sampling in conventional losses [Isola *et al.*, 2017], such as the mean angular loss. This remains a challenging problem to develop an efficient photometric stereo method, which can accurately handle these complex-structured regions.

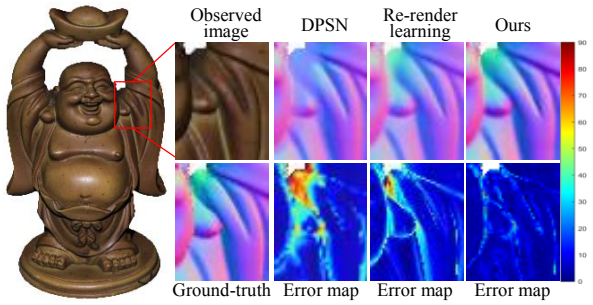

Figure 1: An example of the errors in complex-structured regions. The results compare our method with the DPSN [Santo *et al.*, 2017] and re-render learning [Taniai and Maehara, 2018] methods.

To achieve this goal, we propose an adaptive attention photometric stereo learning framework, called Attention-PSN, which put more emphasis on those areas with high-frequency information. Our framework can significantly reduce the errors that are caused by the unchanged conventional penalty. As shown in Figure 2, Attention-PSN is composed of two parts, which are the normal recovery network and attention network, respectively. The first part of Attention-PSN takes multiple images and illumination directions as the input, and then estimates the surface normals. The second part of Attention-PSN generates an adaptive attention map from the corresponding input images. Then, the adaptive attention map provides the weights for the pixel-wise attention-weighted loss. The adaptive attention map is learned in a self-supervised way, by minimizing the attention-weighted loss. The attention-weighted loss is composed of the angular loss and the gradient loss. A pixel, which has a large value in the attention map, should be paid with more attention anf should have a high detail-preserving level. Consequently, a higher

*Contact Author

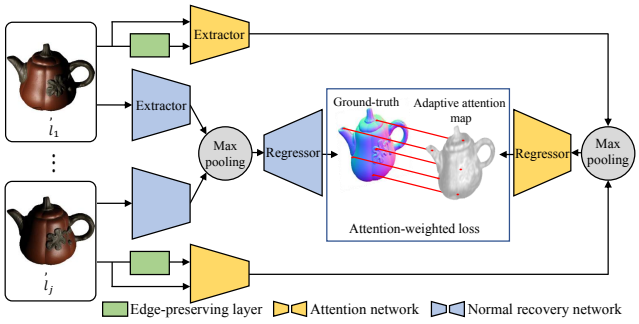

Figure 2: An overview of Attention-PSN. The adaptive attention map provides the weights for the pixel-wise attention-weighted loss. The surface normal and the corresponding adaptive attention map are learned jointly by minimizing the attention-weighted loss.

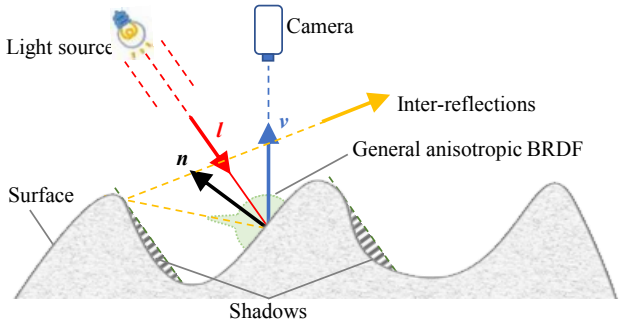

Figure 3: An example of a surface with complex structure, where a surface point with a normal vector $\boldsymbol{n}$ is illuminated by the light $\boldsymbol{l}$, and observed by a camera in a direction $\boldsymbol{v}$. In this instance, shadows and inter-reflections exist widely, accompanied by a general non-Lambertian BRDF (where the green example can be seen as the intensity of reflection in different directions).

weight on the gradient loss and a higher penalty on the high-frequency information should be applied. In these complex-structured regions, the conventional mean angle error loss and Euclidean loss may bring more blurred results [Isola *et al.*, 2017]. In these complex-structured areas, our Attention-PSN learns the pattern from the attention-weighted loss, maintaining the completeness of the high-frequency information. Therefore, Attention-PSN outperforms existing state-of-the-art methods. Extensive experiments on public real datasets show that Attention-PSN achieve promising state-of-the-art results, compared with existing approaches.

2 Related Work

In this section, we briefly review photometric stereo algorithms and learning-based photometric stereo methods. Please refer to [Shi *et al.*, 2019] for a more comprehensive survey of photometric stereo algorithms.

Conventional photometric stereo [Woodham, 1980] suffers from the limitations of the Lambertian reflectance model, failing to meet the need of real-world complex objects. Therefore, a variety of methods have been proposed to handle non-Lambertian surfaces, such as expectation-maximization [Wu and Tang, 2009], rank minimization [Wu *et al.*, 2010], and sophisticated analytical models [Chung and Jia, 2008; Ruiters and Klein, 2009; Holroyd *et al.*, 2008]. However, these approaches can only handle limited classes of non-Lambertian surfaces and cost numerous computations.

Recently, several deep learning methods have been introduced to surface normal recovery [Ju *et al.*, 2019; Chen *et al.*, 2018]. Santo *et al.* [Santo *et al.*, 2017] first proposed a fully-connected network to learn the surface normal by using photometric stereo images, whereas it fails to utilize the adjacent information embedded in images, which leads to unsatisfactory errors, especially in areas with complex structures. Afterwards, Taniai and Maehara [Taniai and Maehara, 2018] applied a self-supervised learning framework to estimate both the surface normal and albedo of an object. Chen *et al.* [Chen *et al.*, 2018; Chen *et al.*, 2019] devoted efforts to handling an arbitrary number of inputs to photometric stereo networks. These subsequent methods are able to perform better surface normal estimation, because they take advantage of the information embedded in the neighborhood by using the convolutional network. However, the above methods fail to satisfacorily handle complex-structured areas, owing to the smooth effect of convolutional layers, as well as the deficiency in the detail-preserving loss. In contrast, in this paper, our proposed Attention-PSN utilizes the adaptive attention map, which determines the weights of the attention-weighted loss, providing suitable penalty strategies for the different areas of a surface with varying complexity.

3 Preliminaries of Photometric Stereo

Before presenting our approach, we recap the theory formulation in photometric stereo, following the common notations. Suppose that a pixel on a reflective surface with a unit normal $\boldsymbol{n} \in \mathbb{R}^3$ is illuminated by the j-th light source with intensity $e_j \in \mathbb{R}$, and the illumination direction $\boldsymbol{l_j} \in \mathbb{R}^3$ in the j-th image, sequentially. When this surface is observed by a linear-response camera in a view direction $\boldsymbol{v} \in \mathbb{R}^3$, the image formation model can be expressed as follows:

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

where m_j represents the measured intensity of a pixel on the surface in the j-th image, ρ is a bidirectional reflectance distribution function (BRDF), $\max\left(\boldsymbol{n}^\top \boldsymbol{l}_j, 0\right)$ accounts for the attached shadows, and ϵ represents the noise and global illumination effect. Facing the non-Lambertian surface existing widely in real-world applications, a general anisotropic BRDF is hard to be solved by using common methods, such as linear least square [Woodham, 1980]. Fortunately, it has been better solved by deep learning methods with stronger fitting capabilities. However, complex-structured areas still suffer relatively larger angle errors. Figure 3 illustrates such a complex-structured area. Therefore, we propose a powerful method that introduces the attention-weighted loss to handle different areas on an object.

4 Proposed Method

In this section, we introduce our method, called Attention-PSN, for better handling calibrated photometric stereo with complex-structured areas and general reflectance. We first

introduce the attention network and the normal recovery network, as shown in Figure 2. Then, we illustrate the attention-weighted loss, the implementation details and datasets.

4.1 Network Architecture

Attention Network

Given a tensor $\boldsymbol{I} \in \mathbb{R}^{JC\times H\times W}$ (where J represents the number of input images, C = 3 is the number of color channels of images, and $H \times W$ is the spatial resolution), we obtain a combined feature map $\Psi \in \mathbb{R}^{JD\times H'\times W'}$ at the end of the extractor (where D is 128 in our network, which is the number of feature maps generated) as:

$$\Psi = f_{\mathrm{ae}}(\boldsymbol{I}, f_{\mathrm{ed}}(\boldsymbol{I}); \theta_{\mathrm{ae}}), \tag{2}$$

where f_{ae} is a three-layer CNN with learnable parameters θ_{ae}. Here, f_{ed} is an edge-preserving layer, which is calculated by the gradient of $\boldsymbol{I}$. Following [Chen *et al.*, 2018], we use max-pooling for multi-feature fusion from $\Psi \in \mathbb{R}^{JD\times H'\times W'}$ to $\Psi' \in \mathbb{R}^{D\times H'\times W'}$, and then output an adaptive attention map P given Ψ', as follows:

$$P = f_{\mathrm{ar}}(\Psi'; \theta_{\mathrm{ar}}), \tag{3}$$

where f_{ar} is a three 3×3 convolutional layers regressor network with learnable parameters θ_{ar}.

Normal Recovery Network

The normal recovery network estimates surface normals of objects. Note that the light direction $\boldsymbol{l}$ is expanded to the same spatial size as the input image, and is concatenated with the image as $\boldsymbol{I}' \in \mathbb{R}^{JC'\times H\times W}$ (where C' = 6 is composed of the RGB channels and the light direction channels.) Then the extractor can be expressed as follows:

$$\Phi = f_{\mathrm{re}}(\boldsymbol{I}'; \theta_{\mathrm{re}}), \tag{4}$$

where f_{re} is the first 118 layers of the ResNet152 [He *et al.*, 2016] with learnable parameters θ_{re}. In particular, we abandon the first pooling layer in ResNet152, considering that excessive pooling layers would lead to a loss of spatial information for the regression task. Similar to attention network, we apply max-pooling to fuse the combined feature $\Phi \in \mathbb{R}^{JD'\times H''\times W''}$ to $\Phi' \in \mathbb{R}^{D'\times H''\times W''}$ (where D' is 512) and then output a surface normal $\bar{\boldsymbol{N}}$, given Φ', as follows:

$$\bar{\boldsymbol{N}} = f_{\mathrm{rr}}(\Phi'; \theta_{\mathrm{rr}}), \tag{5}$$

where f_{rr} is a four 3×3 convolutional layers regressor network with learnable parameters θ_{rr}, ending with an L2 normalization that makes each pixel's normal $\bar{\boldsymbol{n}}$ a unit vector.

4.2 Attention Weighted Loss

We optimize the network parameters θ by minimizing an attention-weighted loss, as follows:

$$\mathcal{L}_{attention} = \frac{1}{HW}\sum_{i}^{HW}\mathcal{L}_i \tag{6}$$

where $\mathcal{L}_i$ is the loss at the pixel i, with resolution $H \times W$, which can be expressed as:

$$\mathcal{L}_i = p_i\mathcal{L}_{\mathrm{gradient}}(\boldsymbol{n}_i, \bar{\boldsymbol{n}}_i) + \lambda(1-p_i)\mathcal{L}_{\mathrm{normal}}(\boldsymbol{n}_i, \bar{\boldsymbol{n}}_i) \tag{7}$$

The first part of the loss, $\mathcal{L}_{\mathrm{gradient}}(\boldsymbol{n}_i, \bar{\boldsymbol{n}}_i)$, defines the gradient loss between the ground-truth $\boldsymbol{n_i}$ and the estimated surface normal $\bar{\boldsymbol{n}}_i$, at pixel i, and is given as follows:

$$\mathcal{L}_{\mathrm{gradient}}(\boldsymbol{n}_i, \bar{\boldsymbol{n}}_i) = \|g(\boldsymbol{n}_{i(x,y)}, \xi) - g(\bar{\boldsymbol{n}}_{i(x,y)}, \xi)\|_2 \tag{8}$$

where (x, y) are the coordinates of i. We define the gradient $g(n_{i(x,y)}, \xi)$ as:

$$\begin{aligned} g(n_{i(x,y)}, \xi) = &\|\frac{f(x+\xi, y) - f(x, y)}{\xi}\|_1 \\ &+\|\frac{f(x, y+\xi) - f(x, y)}{\xi}\|_1 \end{aligned} \tag{9}$$

In our method, ξ is set to 1. Gradient loss can sharpen the discontinuous normal surfaces and prevent these areas from being blurred [Ummenhofer *et al.*, 2017]. We utilize the gradient loss to constrain the completeness of the high-frequency information. However, applying the same gradient loss without using adaptive weights will result in larger errors. This is due to the consequence of suppressing the penalty from other losses in smooth, flat areas (see details in Section 5.1).

The second loss $\mathcal{L}_{\mathrm{normal}}(\boldsymbol{n}_i, \bar{\boldsymbol{n}}_i)$ is a commonly used cosine similarity loss, which directly optimizes the angular error between the ground-truth $\boldsymbol{n}_i$ and the estimated surface normal $\bar{\boldsymbol{n}}_i$. We define $\mathcal{L}_{\mathrm{normal}}(\boldsymbol{n}_i, \bar{\boldsymbol{n}}_i)$ as follows:

$$\mathcal{L}_{\mathrm{normal}}(\boldsymbol{n}_i, \bar{\boldsymbol{n}}_i) = 1 - \boldsymbol{n}_i^{\top}\bar{\boldsymbol{n}}_i \tag{10}$$

As is illustrated in Eq.(7), p_i is the value of the adaptive attention map at the pixel i. λ is a protective threshold, which is set to 8 in our experiments and aims to prevent inadequate penalty on surface normals. By minimizing the attention weighted loss, we learn a self-supervised pattern for different regions and brings the smallest angular error.

4.3 Implementation Details and Datasets

Our network is implemented in PyTorch and the Adam optimizer is used with default settings, where the learning rate is initially set to 0.001 and divided by 2 every 5 epochs. We train the model using a batch size of 64 for 40 epochs and choose a fixed number (j = 16) of images as input, whereas the model accepts an arbitrary number of images in testing.

In our experiments, we use the MERL dataset [Matusik *et al.*, 2003] to render the synthetic 3D model Blobby and Sculpture datasets for training [Johnson and Adelson, 2011]. The MERL dataset contains 100 different BRDFs of real-world materials. Following the settings[Chen *et al.*, 2018], for each sample, 64 images are rendered by random light directions in a half-sphere. We randomly split these samples into a ratio of 99 : 1, for training (84360) and validation (852). In testing, we apply the DiLiGenT dataset [Shi *et al.*, 2019] and the Light Stage Data Gallery [Einarsson *et al.*, 2006].

5 Experimental Results

We perform network analysis for our method and compare our method with the existing state-of-the-art methods on real datasets. To evaluate the accuracy of the estimated surface normals, we adopt the widely used mean angular error (MAE) in degree, denoted as $mean(arccos(\boldsymbol{n}_i \cdot \bar{\boldsymbol{n}}_i))$. Furthermore,

Variants	MAE	err_{15°	err_{30°
Normal loss only	13.10	81.25%	92.32%
Gradient loss only	82.41	0.33%	2.93%
Normal + Gradient	15.48	80.91%	92.80
Attention-weighted loss	**11.77**	**83.07%**	**93.49%**

Table 1: Comparison of the different losses. The numbers represent the average MAE on all samples in the validation set in degrees (the lower the better). The percentages represent the ratio of pixels with the angular error of less than 15° or 30° (the higher the better).

we apply err_{15° and err_{30°, which is the ratio of angular error less than 15° and 30°, respectively. MAE measures the average error of all pixels in an image, while err_{15° and err_{30° indicate the ability of a method to handle complex-structured regions with large errors.

5.1 Network Analysis

We quantitatively analyze Attention-PSN using the MAE, err_{15°, and err_{30° metrics, based on all the samples in the validation set (using 64 input images). We evaluate the effectiveness of the attention-weighted loss in surface normal recovery by comparing it with fixed combination losses as well as the conventional loss. For the conditions without using attention-weighted loss, we only use the normal recovery network. Results are summarized in Table 1.

As shown in Table 1, the attention-weighted loss consistently performs better than the others in all the metrics. It can also be seen that using the gradient loss only fails to make the network converge. Interestingly, in terms of the fixed loss combining both the gradient and normal, we find that the MAE is worse, but err_{30° is improved compared with using the normal loss only.

Discussion on the Attention-weighted Loss

In Table 1, we compare the results based on four types of loss. Now, we discuss the effect of the attention weighted loss.

Firstly, we find that the combined use of both the normal loss and the gradient loss improve err_{30°, compared with using the normal loss only. The gradient loss is activated when there is a large, discontinuous difference between adjacent pixels. It is known that larger angular error mainly exists along edges and the complex-structured regions of objects. Therefore, gradient loss can bring a better constraint on these regions. However, it also shows a worse MAE, which measures the accuracy of all the pixels in an image, compared to using normal loss only.

Furthermore, we witness a non-convergence of the network by using the gradient loss only. We infer that the gradient loss only provides the difference between adjacent pixels, ignoring the pixel's value. This can also explain why MAE becomes worse in mixed losses, because the gradient loss dilutes the penalty in flat areas.

As for attention-weighted loss, a higher err_{15° and err_{30° mean that fewer pixels suffer from the large angular error. In these complex-structured areas, Attention-PSN learns the pattern from the attention-weighted loss, maintaining the completeness of the high-frequency information. Meanwhile, attention-weighted loss also performs the best in terms of MAE. This is because Attention-PSN learns the pattern from lower weights of the gradient loss on flat (low-frequency) regions, avoiding the adverse impact to the surface normal.

5.2 DiLiGenT Benchmark Comparisons

The DiLiGenT benchmark [Shi *et al.*, 2019] contains ten real-world scenes of photometric stereo, which is challenging for its strong non-Lambertian surfaces and complex structures. We show our results on the DiLiGenT benchmark in Table 2, where we compare our method with existing methods in terms of MAE. We also show the estimations and attention maps, compared with learning-based approaches in Figure 4, including DPSN [Santo *et al.*, 2017], PS-FCN [Chen *et al.*, 2018], Re-render Learning [Taniai and Maehara, 2018], the best non-learning-based method ST14 [Shi *et al.*, 2014], as well as the baseline using least square [Woodham, 1980]. In this experiment, we use 96 observed images for each scene provided by the DiLiGenT dataset as an input.

Table 2 and Figure 4 compare the normal estimation results of Attention-PSN with previous state-of-the-art calibrated photometric stereo methods on the DiLiGenT benchmark. Attention-PSN achieves state-of-the-art results on most of the objects, with an average MAE of 7.92 (trained with 16 images per-sample and tested with all the 96 images for each object). As for the object "Bear", our method significantly reduces MAE by 16.1%, compared to the sub-optimal method [Taniai and Maehara, 2018].

Discussion on Benchmark Comparisons

We show some examples in Figure 4 with red boxes, such as the waistband of "Buddha", as well as the flower of "Pot2". It can be seen that attention maps are activated by these complicated areas, where the values are higher. Accordingly, the error maps of our method show a lower angular error in these regions, compared with others. In these regions, the weights for gradient loss in the attention-weighted loss are larger, it indicates that Attention-PSN have learned the pattern of the completeness of high-frequency information. In this way, the estimated normal will keep clear edges with less blur. In contrast, the blur can be easily found in Re-render learning [Taniai and Maehara, 2018], PS-FCN [Chen *et al.*, 2018], and DPSN [Santo *et al.*, 2017]. In these methods, only the single angular loss is used, whereas it performs poorly in the regions with sharp surface normal. This is because the sampling by conventional losses smooths the high-frequency information [Isola *et al.*, 2017]. Furthermore, the excessive receptive filed of deep convolutional neural networks aggravates blur, such as Re-render learning [Taniai and Maehara, 2018] and PS-FCN[Chen *et al.*, 2018]. Instead, we match the detail-preserving gradient loss with the deeper ResNet152 (first 118 layers) extractor through the attention-weighted loss. Thus, Attention-PSN achieves lower MAE while maintaining good high-frequency information.

We also notice that the proposed method does not achieve the best performance on the object "Ball", which is a particularly simple structure. As shown in the yellow boxes in Figure 5, It can be seen that the attention map of Ball is less convincing. The reasonfor this may be that the specular misleads the attention network in handling very simple

Method	Avg.	bear	buddha	goblet	harvest	pot2	pot1	cat	cow	reading	ball
L2 (Baseline)	15.39	8.39	14.92	18.50	30.62	14.65	8.89	8.41	25.60	19.80	4.10
IW12	13.74	7.32	11.11	16.25	29.26	14.09	7.74	7.21	25.70	16.17	2.54
WG10	13.35	6.50	10.91	15.70	30.01	13.12	7.18	6.73	25.89	15.39	2.06
AZ08	12.61	5.96	12.54	13.93	30.50	11.03	7.23	6.53	21.48	14.17	2.71
IA14	10.60	7.11	10.47	9.71	25.95	8.77	6.64	6.74	13.05	14.19	3.34
ST14	10.30	6.12	10.60	10.09	25.44	8.78	6.51	6.12	13.93	13.63	1.74
DPSN	9.41	6.31	12.68	11.28	16.86	7.86	7.05	6.54	8.01	15.51	2.02
Re-render Learning	8.83	5.79	10.36	11.47	22.59	7.76	**6.09**	**5.44**	**6.32**	**11.03**	**1.47**
PS-FCN	8.39	7.55	7.91	8.60	15.85	7.25	7.13	6.16	7.33	13.33	2.82
Attention-PSN (Proposed)	**7.92**	**4.86**	**7.75**	**8.42**	**15.44**	**6.97**	6.92	6.14	6.86	12.90	2.93

Table 2: Comparison of different methods on the DiLiGenT benchmark. All methods are evaluated with 96 images. Here, we measure MAE in degrees. The methods listed are L2 (Baseline) [Woodham, 1980], IW12 [Ikehata *et al.*, 2012], WG10 [Wu *et al.*, 2010], AZ08 [Alldrin *et al.*, 2008], IA14 [Ikehata and Aizawa, 2014] ST14 [Shi *et al.*, 2014], DPSN [Santo *et al.*, 2017], PS-FCN [Chen *et al.*, 2018], and Re-render Learning [Taniai and Maehara, 2018].

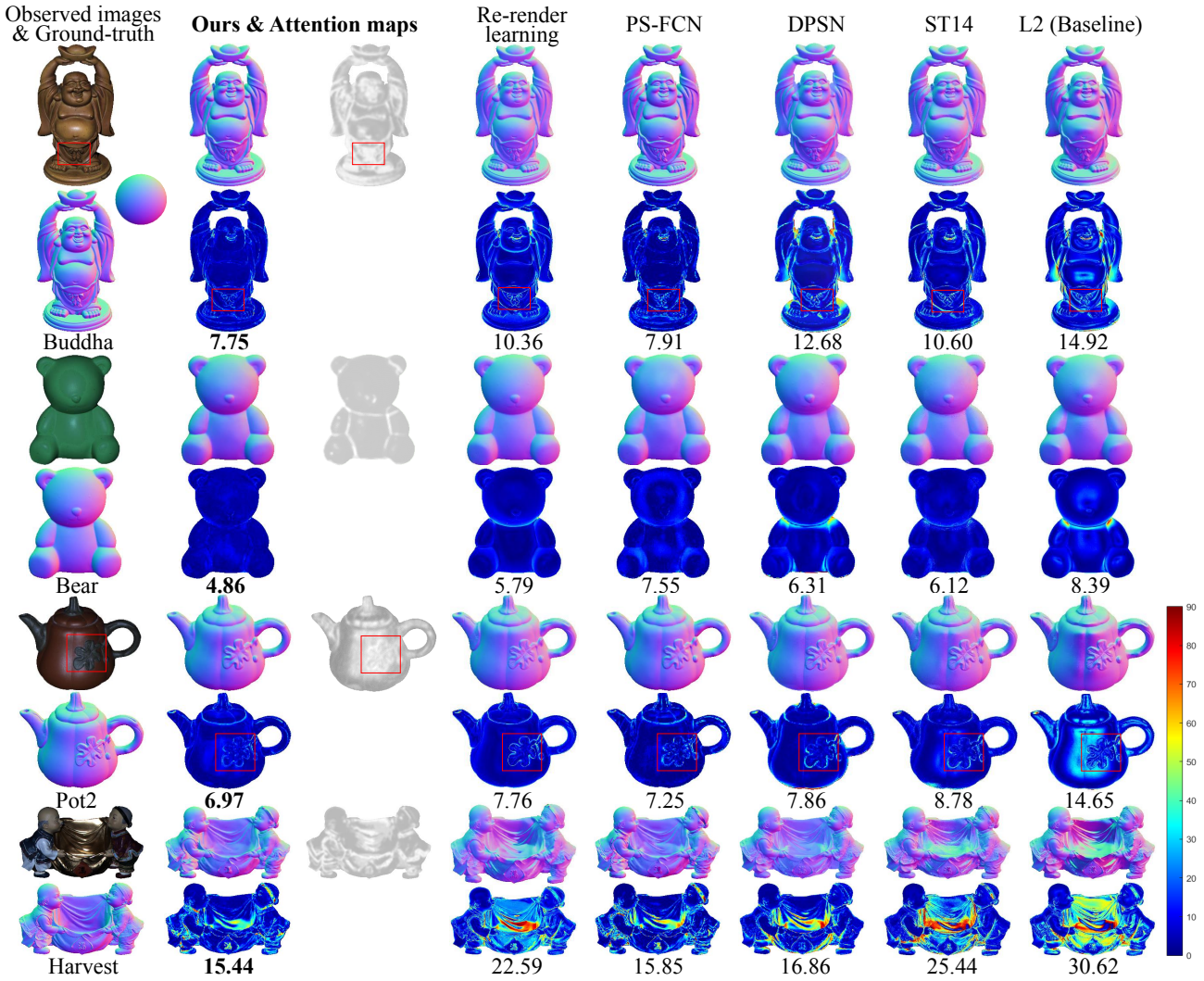

Figure 4: Visual comparisons for Buddha, Bear, Pot2, and Harvest scenes. The red boxes are regions with complex structures (high-frequency information). We adjust the contrast of observed images for easy viewing. From left to right columns in each scene, we show 1) observed images and ground-truth, 2) estimated surface normals and angular error maps by our method, 3) attention maps of our method, and 4-8) estimated surface normal and angular error maps by four state-of-the-art methods and baseline. Numbers under angular error maps show their MAE in degree.

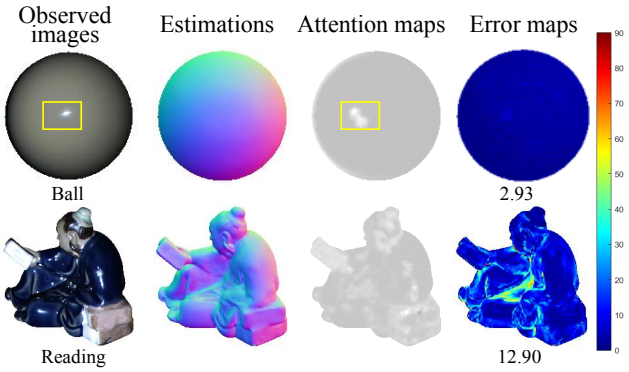

Figure 5: Our results on Ball and Reading. The yellow boxes are regions of specular. We adjust the contrast of observed images for easy viewing. Numbers under Error maps show their MAE in degree.

structures, where the specular is the only high-frequency information. Moreover, our method obtains sub-optimal performance on "Reading" and "Cow", where the structures are not very complicated, but existing strong non-Lambertian and intricate BRDFs. It means that the normal maps (ground-truth) are not as high-frequency as the observed images. Therefore, we infer that the mismatch of the level of the high-frequency information between ground-truths and observed images may impact our results to some extent.

5.3 Testing on the Light Stage Data Gallery

Figure 6 shows the results of Attention-PSN using the Light Stage Data Gallery [Einarsson *et al.*, 2006]. We qualitatively evaluate Attention-PSN on this dataset to further demonstrate the transferability of our method. Due to the absence of ground-truth, we qualitatively show our performance. Light Stage Data Gallery consists of six objects, and 253 images and corresponding light directions, as well as intensities, provided for each object. In this experiment, we choose $j = 144$ as the number of input images.

Discussion on the Light Stage Data Gallery

As shown in Figure 6, the estimations reflect the shapes of the objects, compared with the calibrated sphere. For instance, the skirt of the object "Knight fighting" is made of lumpy-looking material. It can be seen that our results also show the corresponding surface in this area, as well as the higher weights in the attention map. Similarly, we prove the accuracy of our method on the sleeve of the Knight, which also illustrate the effectiveness of our method.

We also observe that the estimated surface normal and attention map of the object "Knight standing" is with some noise. It might be due to the poor quality of the observed images. The Charge Coupled Device (CCD) of the camera can not suppress the noise in the dark environment (higher photo-sensitivity). Thus, the high-frequency noise existing in input images may impact the performance of Attention-PSN.

6 Conclusions

We proposed a deep learning framework, called Attention-PSN, for photometric stereo. Ablation experiments have illustrated that the attention-weighted loss results in higher accurate reconstruction, especially in areas with complex structures. Extensive experiments on the public DiLiGenT benchmark have shown that Attention-PSN outperforms state-of-the-art methods in calibrated photometric stereo. We achieved an average MAE of 7.92 on the DiLiGenT benchmark. Also, visual comparisons have shown the ability of our method in handling complex-structured areas, where our method can achieve the best estimation and reconstruction at high-frequency regions, with minimum blur. Our method obtains promising results with clear details. This demonstrates the robustness of Attention-PSN.

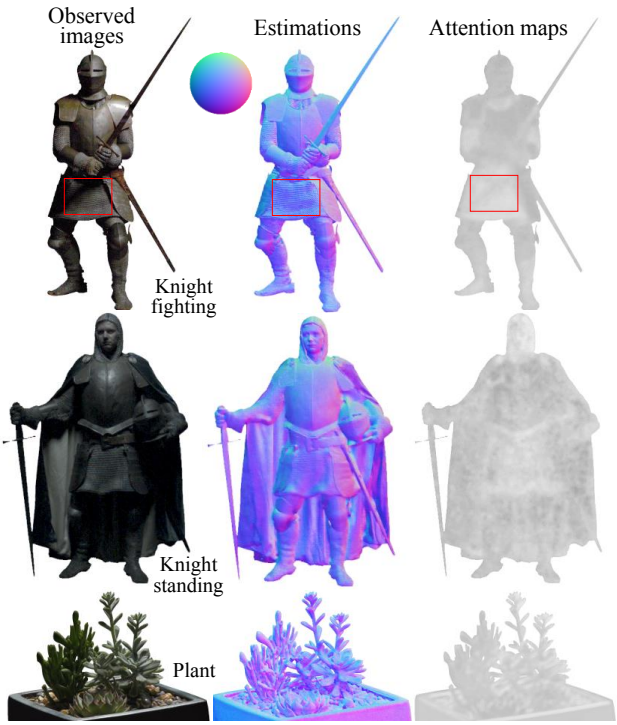

Figure 6: Qualitative results of Attention-PSN for the Light Stage Data Gallery. The red boxes are regions with high-frequency information. We adjust the contrast of observed images for easy viewing.

Furthermore, the proposed attention-weighted loss can also provide a framework for other regression tasks, such as depth estimation and image enhancement. In these tasks, the attention weighted loss can learn an adaptive penalty , and recover a clear estimation with less blur.

Acknowledgments

The work was supported by the National Key R & D Program of China under Grant (2018AAA0100602), the National Key Scientific Instrument and Equipment Development Projects of China (41927805), the National Natural Science Foundation of China (61501417, 61976123), and the Joint Funds of the National Natural Science Foundation of China-Shandong (U1706218). We thank Guanying Chen for code and help. We also thank Hiroaki Santo for his help with the providing of comparison results.

References

[Alldrin *et al.*, 2008] 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*, pages 1–8. IEEE, 2008.

[Chen *et al.*, 2018] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[Chen *et al.*, 2019] 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*, pages 8739–8747, 2019.

[Chung and Jia, 2008] Hin-Shun Chung and Jiaya Jia. Efficient photometric stereo on glossy surfaces with wide specular lobes. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[Einarsson *et al.*, 2006] Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. Relighting human locomotion with flowed reflectance fields. In *Proceedings of the 17th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association, 2006.

[He *et al.*, 2016] 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*, pages 770–778, 2016.

[Holroyd *et al.*, 2008] Michael Holroyd, Jason Lawrence, Greg Humphreys, and Todd Zickler. A photometric approach for estimating normals and tangents. In *ACM Transactions on Graphics (TOG)*, volume 27, page 133. ACM, 2008.

[Ikehata and Aizawa, 2014] Satoshi Ikehata and Kiyoharu Aizawa. Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2186, 2014.

[Ikehata *et al.*, 2012] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa. Robust photometric stereo using sparse regression. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 318–325. IEEE, 2012.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[Jian *et al.*, 2019] Muwei Jian, Junyu Dong, Maoguo Gong, Hui Yu, Liqiang Nie, Yilong Yin, and Kin-Man Lam. Learning the traditional art of chinese calligraphy via three-dimensional reconstruction and assessment. *IEEE Transactions on Multimedia*, 2019.

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

[Ju *et al.*, 2019] Yakun Ju, Xinghui Dong, Yingyu Wang, Lin Qi, and Junyu Dong. A dual-cue network for multispectral photometric stereo. *Pattern Recognition*, page 107162, 2019.

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

[Ruiters and Klein, 2009] Roland Ruiters and Reinhard Klein. Heightfield and spatially varying brdf reconstruction for materials with interreflections. In *Computer Graphics Forum*, volume 28, pages 513–522. Wiley Online Library, 2009.

[Santo *et al.*, 2017] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 501–509, 2017.

[Shi *et al.*, 2014] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1078–1091, 2014.

[Shi *et al.*, 2019] B Shi, Z Mo, Z Wu, D Duan, SK Yeung, and P Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):271–284, 2019.

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

[Ummenhofer *et al.*, 2017] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047, 2017.

[Woodham, 1980] R. J Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[Wu and Tang, 2009] Tai-Pang Wu and Chi-Keung Tang. Photometric stereo via expectation maximization. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):546–560, 2009.

[Wu *et al.*, 2010] 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*, pages 703–717. Springer, 2010.