Pattern Recognition Letters
journal homepage: www.elsevier.com

# Deep Uncalibrated Photometric Stereo via Inter-Intra Image Feature Fusion

Fangzhou Gao[a], Meng Wang[a], Lianghao Zhang[a], Li Wang[a], Jiawan Zhang[a,**]

[a]*Department of Intelligence and Computing, Tianjin University, Tianjin, China*

*Article history*:

*2000 MSC:* 41A05, 41A10, 65D05, 65D17

*Keywords:* uncalibrated photometric stereo, deep neural network

ABSTRACT

Uncalibrated photometric stereo is proposed to estimate the detailed surface normal from images under varying and unknown lightings. Recently, deep learning brings powerful data priors to this underdetermined problem. This paper presents a new method for deep uncalibrated photometric stereo, which efficiently utilizes the inter-image representation to guide the normal estimation. Previous methods use optimization-based neural inverse rendering or a single size-independent pooling layer to deal with multiple inputs, which are inefficient for utilizing information among input images. Given multi-images under different lighting, we consider the intra-image and inter-image variations highly correlated. Motivated by the correlated variations, we designed an inter-intra image feature fusion module to introduce the inter-image representation into the per-image feature extraction. The extra representation is used to guide the per-image feature extraction and eliminate the ambiguity in normal estimation. We demonstrate the effect of our design on a wide range of samples, especially on dark materials. Our method produces significantly better results than the state-of-the-art methods on both synthetic and real data.

## 1. Introduction

Photometric stereo is proposed to estimate the surface normal from images captured by a fixed camera under varying and known lighting. Compared with other stereo vision methods like multi-view stereo, photometric stereo can produce more detailed normal and perform well on textureless objects. The pioneering photometric stereo method is proposed for the idea Lambertian surface [1], following researchers have extended it to handle a wide range of complex surfaces [2–13].

However, these methods rely on complex light calibration. To overcome it, uncalibrated photometric stereo is proposed to accomplish the task without light calibration. To reduce the ill-posedness caused by the lack of lighting information, most traditional methods assume an ideal Lambertian surface [14–16] or a uniform light distribution [17, 18], which limits the practical application. Recently, motivated the significant advancements made by deep learning in computer vision, some researchers [19–22] have utilized deep learning to leverage data priors and generalized the uncalibrated photometric stereo to the real complex condition.

The main challenge in deep uncalibrated photometric stereo is to enable the network to deal with the unordered and arbitrary numbers of input images. The common CNN-based network is unsuitable since it requires fixed input channels. Some researchers utilized optimization-based neural inverse rendering to solve this problem [22]. Kaya *et al.* [22] optimized the surface normal by the neural rendering layers. They explicitly modeled the effect of interreflection and did not rely on the ground-truth of surface normals for training. While their method was limited by the assumption of the differentiable surface and performed poorly on complicated objects. Other researchers used the size-independent pooling layer to aggregate features from different inputs [20, 21]. Chen *et al.* [20] proposed a light calibration network to estimation the lighting for the following normal estimation. They first used a shared-weight feature extractor to explore intra-image variation from each input independently and then fused them using a max-

**Corresponding author.
*e-mail:* gaofangzhou@tju.edu.com (Fangzhou Gao), meng.wang@tju.edu.cn (Meng Wang), opoiiuiouiuy@tju.edu.cn (Lianghao Zhang), li_wang@tju.edu.cn (Li Wang), jwzhang@tju.edu.cn (Jiawan Zhang )

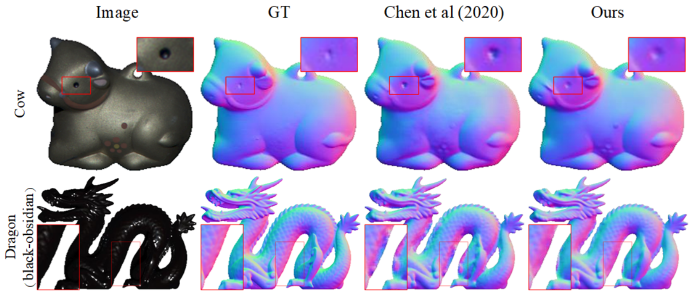

**Fig. 1. Visualized results of the cow sample in [23] and the dragon sample in [21], compared with the state-of-the-art method [21]. Chen *et al.* [21] misjudges the normals on dark materials, especially on regions that lack highlights. In contrast, our method significantly improves the results. The dark regions are marked with red boxes and enlarged.**

pooling layer to explore the inter-image variation. Chen *et al.* [21] further designed a cyclic network structure to introduce extra inter-image guidance and intra-image guidance for the lighting estimation. It improved the accuracy in lighting estimation. However, the single pooling layer was weak in exploring input images' information, which caused the estimated surface normals were still not satisfactory, especially on dark materials, as shown in Fig. 1.

In this paper, we consider the correlation between intra-image intensity variation and inter-image lighting variation and propose an inter-intra image feature fusion module to combine these two kinds of variations. Specifically, we explore the intra-image variation from each input by a share-weight CNN-based feature extractor that contains our feature fusion modules. During per-image local feature extraction, these fusion modules aggregate global features (like material and rough geometry) among multi-images and introduce them into local feature extraction. The implicit material and geometry representations in global features can guide the next local feature extraction, which allows a more efficient feature extraction and a more accurate normal estimation. Experiments demonstrate that our design significantly improved the results, especially on the challenging dark materials.

## 2. Related Work

### 2.1. *Deep Uncalibrated Photometric Stereo*

Most traditional methods in uncalibrated photometric stereo rely on unpractical assumptions to like an idea Lambertian reflectance model [1, 24, 25] or a uniform light distribution [26, 27]. On the contrast, the learning-based methods leveraged powerful data priors and performed better on real objects.

To enable the network handle arbitrary numbers of input images in uncalibrated photometric stereo, some researchers utilized neural inverse rendering [22], while others utilized size-independent pooling layers to fuse features.

Kaya *et al.* [22] calculated the surface normals, BRDFs, and depth by the optimization of neural rendering loss, which explicitly modeled the interreflections. While their neural rendering relied on a continuous surface to compute depth and

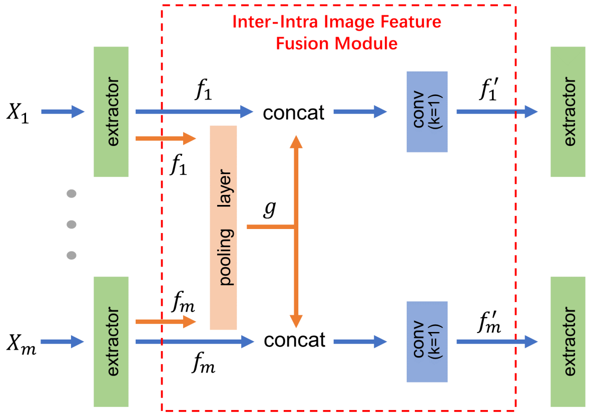

**Fig. 2. An illustration of the inter-intra image feature fusion module. It aggregates global features from the intermediate local features of each input and concatenates them with per-image local features separately for the following 1 × 1 convolutional layer, which fuses local features with global features.**

interreflection kernel, and performed poorly on complex surfaces. Chen *et al.* [19] directly predicted surface normal from input images. They used a shared-weight extractor to extract local features from each input before fusing them using a max-pooling layers. Chen *et al.* [20] further introduced lightings as extra supervision. They first estimated the light directions and intensities, then predicted the surface normral with estimated lightings. Recently, Chen *et al.* [21] designed a cyclic network structure for lighting estimation. They first estimated rough lightings and rough surface normal, then provided computed shading and rough normal as extra guidance for the final lighting estimation. But they only focus on improving accuarcy in lighting estimation, the final results of surface normals need to further improved.

### 2.2. *Multi-Image Deep Network*

Similar with photometric stereo, many other tasks in computer vision and computer graph take a variable number of images as input [28–32]. Choy *et al.* [28] took images as a squeeze and applied a RNN-based network in multi-view 3D reconstruction. While their architecture was sensitive to the order of inputs and paid less attention to latter images. To overcome it, Wiles *et al.* [29] used a shared-weight feature extractor to extract local features from each image, then fused them to a fix-sized global feature using a order-independent pooling layer. Similar strategies were also adopted in SVBRDF capture [30], burst image deblurring [31], deep learning on 3d points [32]. Inspired by these works, we designed our network for uncalibrated photometric stereo, which leverages the correlation between images efficiently.

## 3. Our Method

This section firstly introduces our motivation and strategy of exploring the intra-image and inter-image variations and then presents our network structure.

Following the common assumptions, we assume that images are captured by a radiometrically calibrated orthogonal camera

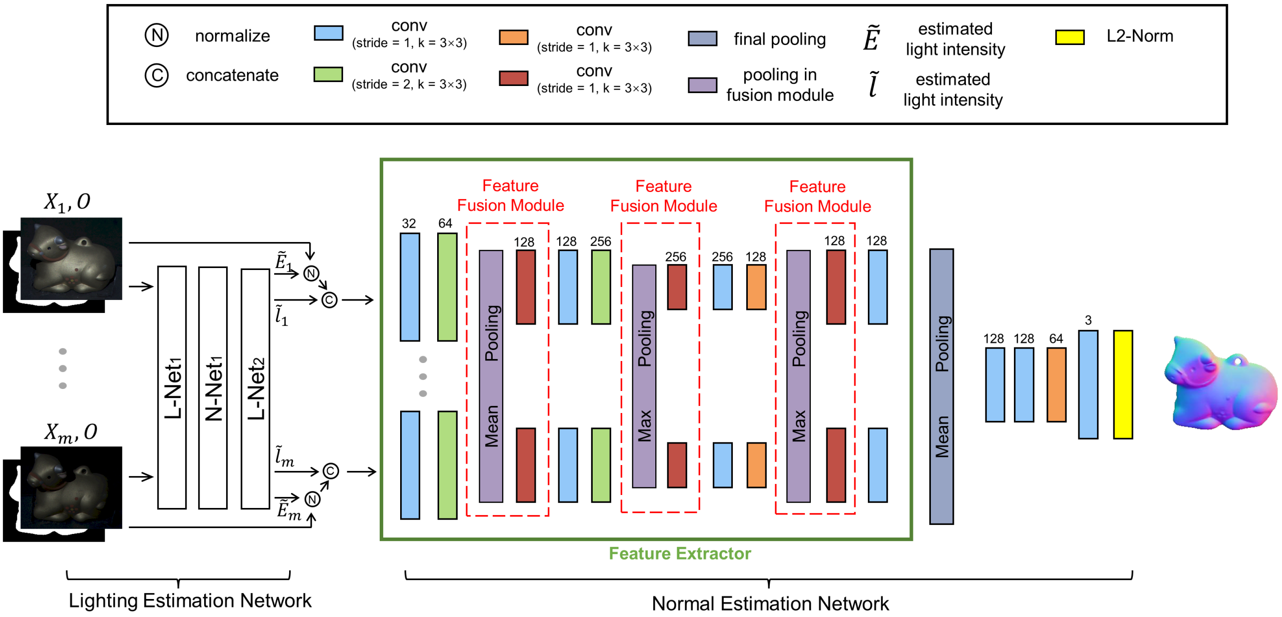

**Fig. 3. Overview of our framework. With our feature fusion module, our normal estimation network combines inter-image features with local features during local feature extraction.**

under single directional lighting. Moreover, we use “intensity” to refer to image irradiance for simplicity.

### *3.1. Inter-Intra Image Feature Fusion Module*

In deep uncalibrated photometric stereo, the intra-image intensity variation and inter-image lighting variation of input images are correlated and significant for normal estimation. The inter-image variation under different light conditions implies the surface material and geometry information, which is crucial for eliminating the ambiguity of the normal, light and reflectance model. A common approach to exploring these two kinds of variation is to use a shared-weight extractor to extract per-image features independently from each input, then fuse them using a pooling layer [19, 29, 30]. However, this network structure can not perceive any inter-image variation to eliminate the ambiguity in per-image feature extraction.

To over this problem, we propose the inter-intra image feature fusion module. Specifically, For a set of input images $X = [X_1, X_2, ..., X_m]$ a set of $M$ input images, there are a set of per-image features $f = [f_1, f_2, ..., f_m]$ extracted by the front layers in the share-weight extractor. The modules inserted fuses features as follow:

$$g = Pool\left(f_1, f_2, \cdots\cdots, f_m\right), \tag{1}$$

where *Pool* represents the pooling layer that aggregates global features $g$ from intermediate local features $f$.

$$f_i' = N_{fuse}\left(f_i, g\right), \tag{2}$$

where $N_{fuse}$ represents the 1×1 convolution layer that fuses local features with global features separately to obtain the new per-image features $f_i'$. The new per-image features $f' = [f_1', f_2', ..., f_m']$ are fed to the rest layers in the extractor, as shown in Figure 2.

With the global features representing the inter-image variation of all images, the extractor can utilize the implied material and geometry features to extract local features more accurately and efficiently. For instance, it is easy for the extractor to roughly distinguish the shadows and regions with low albedo since the intensities of common shadows change rapidly with the changing lighting while the intensities of regions with low albedo remain low. Moreover, with the extra cues of other inputs, the network can capture the slight changes of the intensities of dark materials, which provide strong cues for inferring the surface normal.

### *3.2. Network Structure*

Our method contains two networks: the lighting estimation network and the normal estimation network. Given $X = [X_1, X_2, ..., X_m]$ a set of $M$ input images and the object mask O, as shown in Figure 3, we first estimate the light directions and intensities for each image using the lighting estimation network. Then the estimated lighting is used to recover the surface normal using our proposed normal estimation network.

**Lighting Estimation Network** For the lighting estimation network, we follow the structure proposed in [21]. As shown in Figure 3, the lighting estimation network contains three sub-networks, including two lighting estimation sub-networks (L-Net$_1$ and L-Net$_2$) and a normal estimation sub-network (N-Net). The L-Net$_1$ estimates initial lighting given the input images and object mask. Then the N-Net predicts surface normal given the initial lighting and input images. Finally, the L-Net$_2$ estimates the final lighting with extra rough normal and shading estimated by the front networks. More details can be found in [21].

**Normal Estimation Network** With the estimated lighting, we normalize input images with the corresponding predicted

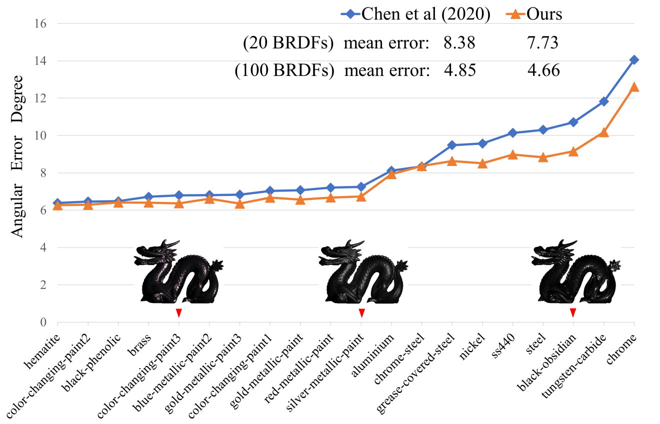

**Fig. 4. Quantitative results of of dragon test data in [21], compared with Chen *et al.* [21]. Our method has lower mean error for both 20 BRDFs shown in the figure and all 100 BRDFs. Besides, our method performs significantly better for those dark materials which are challenging for Chen's method. Images above the horizontal axis show the corresponding samples**

light intensity, and then concatenate them with the corresponding predicted light direction as the inputs. As shown in Figure 3, for a set of image-lighting pairs, per-image local features are extracted separately by a shared-weight feature extractor before fusing them to global features. Then the following several convolutional layers and an L2-normalization layer further infer the normal map from global features.

The CNN-based feature extractor is plugged with the inter-intra image feature fusion modules at multiple levels. During the local feature extraction, as motioned in section 3.1, the feature fusion modules introduce the global features among multi-images to guide the per-image feature extraction. We choose the mean-pooling layer for the first fusion block to obtain the material and reflectance representations and the max-pooling layer for others.

### 3.3. Loss Function

The lightings are discretized and considered as a classification problem (32 classes for elevation and azimuth to represent light direction, 32 classes for intensity). Given $M$ images, the loss function for L-Net in lighting estimation network is

$$\mathcal{L}_{\text{light}} = \frac{1}{M}\sum_{f}\left(\mathcal{L}^m_{l_a} + \mathcal{L}^m_{l_e} + \mathcal{L}^m_{e}\right), \tag{3}$$

where $\mathcal{L}^m_{l_a}$,$\mathcal{L}^m_{l_e}$ and $\mathcal{L}^m_{e}$ are the cross-entropy loss for light azimuth, elevation, and intensity classifications, respectively.

The normal loss function of the normal estimation network and the N-Net in lighting estimation network is

$$\mathcal{L}_{\text{normal}} = \frac{1}{P}\sum_{p}\left(1 - \boldsymbol{n}_p^{\top}\tilde{\boldsymbol{n}}_p\right), \tag{4}$$

where $P$ donates the number of pixels in per image, and $\boldsymbol{n}_p$ and $\tilde{\boldsymbol{n}}_p$ are the ground truth and predicted normal at pixel $p$, respectively. And we fine-tune the entire lighting estimation network end-to-end using the following loss:

$$\mathcal{L}_{\text{fine-tune}} = \mathcal{L}_{\text{light}_1} + \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{shading}} + \mathcal{L}_{\text{light}_2}, \tag{5}$$

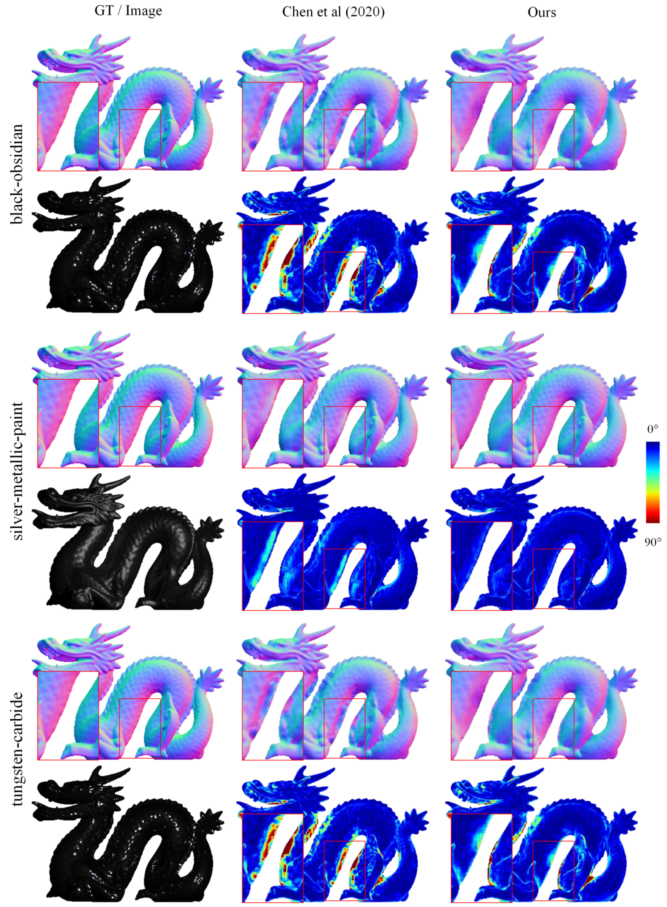

**Fig. 5. Visualized results of dragon test data in [21], compared with Chen *et al.* [21]. The visualized error maps lie underneath the estimated normal maps. The regions where lacks highlight due to self-occlusion are marked with red boxes and enlarged at the bottom-left corner.**

$$\mathcal{L}_{\text{shading}} = \frac{1}{MP}\sum_{m}\sum_{p}\left(\boldsymbol{n}_p^{\top}\boldsymbol{l}_m - \tilde{\boldsymbol{n}}_p^{\top}\tilde{\boldsymbol{l}}_m\right)^2, \tag{6}$$

where $\mathcal{L}_{\text{shading}}$ denote $\mathcal{L}_{\text{light}_1}$ and $\mathcal{L}_{\text{light}_2}$ denote the loss function in L-Net$_1$ and L-Net$_2$, $\mathcal{L}_{\text{shading}}$ denote the shading loss, and $\boldsymbol{l}_m$ and $\tilde{\boldsymbol{l}}_m$ are the ground truth and predicted light direction for the $m^{th}$ image.

## 4. Experiment

We evaluated and analyzed our method on synthetic and real data and used the popular mean angular error (MAE) to measure the error of the estimated normal.

### 4.1. implement details

We trained our model on the publicly available synthetic Blobby and Sculpture Dataset [19], which contains 85,212 surfaces and each is illuminated under 64 random light directions.

For the lighting estimation, we followed the training procedure in [21] to train three sub-networks one after another until convergence, then fine-tuned the lighting estimation network

**Table 1. Quantitative results on DiLiGenT benchmark [23], compared with comparison with traditional and deep uncalibrated photometric stereo methods. For each object, the best result is bolded and colored in dark-red, and the second best result is colored in light-red. † means we use the deeper vision of version of UPS-FCN. [19]**

| Methods | Ball | Cat | Pot1 | Bear | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Papadh. et al. (2014) [15] | 4.77 | 9.54 | 9.51 | 9.07 | 15.90 | 14.92 | 29.93 | 24.18 | 19.53 | 29.21 | 16.66 |
| Lu et al. (2017) [33] | 9.30 | 12.60 | 12.40 | 10.90 | 15.70 | 19.00 | 18.30 | 22.30 | 15.00 | 28.00 | 16.30 |
| UPS-FCN† (2018) [19] | 3.96 | 12.16 | 11.13 | 7.19 | 11.11 | 13.06 | 18.07 | 20.46 | 11.84 | 27.22 | 13.62 |
| SDPS-Net (2019) [20] | 2.77 | 8.06 | 8.14 | 6.89 | 7.50 | 8.97 | 11.91 | 14.90 | 8.48 | 17.43 | 9.51 |
| Chen et al.(2020) [21] | **2.48** | 7.87 | **7.21** | 5.55 | **7.05** | 8.58 | 9.62 | 14.92 | 7.81 | 16.22 | 8.73 |
| Kaya et al. (2021) [22] | 3.78 | 7.91 | 8.75 | 5.96 | 10.17 | 13.14 | 11.94 | 18.22 | 10.85 | 25.49 | 11.62 |
| Ours | 3.04 | **7.55** | 7.54 | **5.40** | 8.05 | **8.39** | **8.91** | **14.81** | **6.88** | **15.23** | **8.58** |

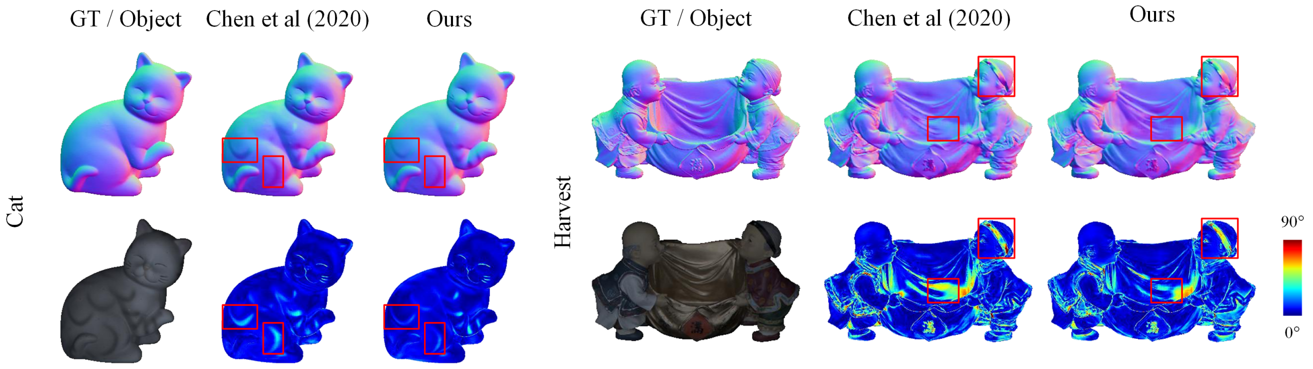

**Fig. 6. Visualized results of DiLiGenT benchmark [23], compared with Chen *et al.* [21]. The object image is produced by averaging all images for better visualization. The visualized error maps lie underneath the estimated normal maps. The red boxes demonstrate our improvement in dark materials and concave regions. The full results of our method on DiLiGenT benchmark are included in the supplementary.**

end-to-end. The normal estimation network was trained with ground truth lighting and surface normal, with a batch size of 32 for 30 epochs. We used the same training configuration in [21], including the learning rate, batch size, number of training epochs, etc, to further demonstrate the superiority of our network architecture.

We implemented the framework in PyTorch and used the Adam optimizer [34] with default parameters. It took 4.63 hours to train the normal estimation network with a 3.70GHz Intel Core i9 CPU and a single NVIDIA GeForce RTX 3060 GPU.

### *4.2. Evaluation on Synthetic Data*

We compared our method with the state-of-the-art method [21] on the synthetic dragon dataset in [21]. The dragon shape is rendered with 100 MERL BRDFs [35] and each is illuminated under 82 randomly sampled light directions.

In Figure 4, we show quantitative results of 20 BRDFs on which Chen's method [21] performed worst and sort them by the error. Our method produced better results, especially for those challenging dark materials, as we analyzed in section 3.1. Figure 5 shows the visualized results of several dark materials. For marginal regions where lack highlights as extra cues, Chen's method produced unreasonable surface normals, while our method significantly improved them on various materials. It proves the superiority of our method. With the feature fusion modules, our model can infer the surface normal through slight intensities changes among images.

### *4.3. Evaluation on Real Data*

We evaluated our method on the public DiLiGenT benchmark [23] and report the quantitative results compared with other uncalibrated photometric stereo methods. As shown in Table 1, our method achieved the best performance with the lowest average error and the lowest error for most objects. The visualized results in Figure 6 and Figure 1 also proves that our method significantly improve the normal estimation on dark and concave regions.

Besides, we demonstrated the superiority of our method on the feature domain. We compared our method with Chen *et al.* [21], in which the extractor dose not perceive any inter-image information. We averaged all channels of the fused global features aggregated by the final pooling layer (each channel was normalized). As shown in Figure 7, few valid features were extracted for dark regions in Chen's method, which explained its poor result of surface normal. While our method extracted valid features for all regions and inferred more accurate normals.

### *4.4. Ablation Study*

To validate the effect of perceiving inter-image variation during local feature extraction, we removed all feature fusion modules, then trained our network from scratch in the same configuration. And we also evaluated the model that only be removed the pooling layers in feature fusion modules to prove our method does not simply benefit from a deeper network. The quantitative results of DiLiGenT benchmark [23] and dragon test data in [21] are summarized in Table 2. The visualized results are shown in Figure 8. Ours $v_1$ represents the model

**Table 2. Quantitative results of the ablation study. For the DiLiGent benchmark, We report the error of all objects and the average error. For Synthetic dragon data, we report the errors of samples of ten typical materials and the average error of all 100 BRDFs. The lowest errors are bolded.**

| Methods | DiLiGenT benchmark | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ball | Cat | Pot1 | Bear | Pot2 | Buddha | Goblet | Reading | Cow | Harvest | Average |
| Ours | 3.04 | **7.55** | **7.54** | 5.40 | **8.05** | **8.39** | **8.91** | **14.81** | **6.88** | **15.23** | **8.58** |
| Ours v$_1$ | **2.85** | 9.1 | 7.86 | **5.35** | 8.75 | 8.84 | 9.42 | 15.63 | 7.56 | 16.45 | 9.18 |
| Ours v$_2$ | 3.7 | 8.61 | 7.63 | 6.17 | 9.20 | 8.44 | 9.45 | 15.11 | 7.61 | 16.23 | 9.21 |

| Methods | Synthetic dragon data | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Aluminium | Blue-acrylic | Chrome | Delrin | Nickel | Nylon | PVC | SS440 | Steel | Tungsten-Carbide | Average |
| Ours | **7.93** | **3.79** | **12.62** | **3.25** | **8.51** | **3.85** | 3.59 | **8.98** | **8.84** | **10.18** | **4.66** |
| Ours v$_1$ | 8.57 | 4.65 | 13.91 | 3.55 | 9.88 | 4.69 | **3.58** | 10.14 | 10.20 | 11.55 | 5.03 |
| Ours v$_2$ | 8.33 | 5.21 | 14.01 | 3.36 | 9.58 | 4.80 | 3.95 | 9.74 | 11.20 | 11.13 | 5.03 |

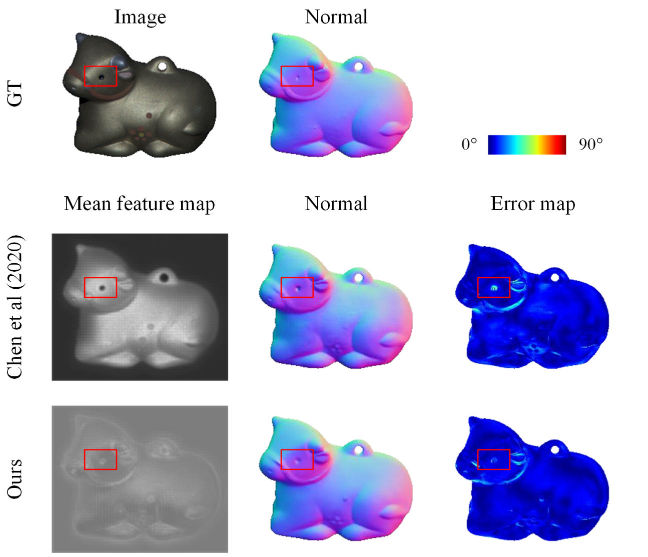

**Fig. 7. Visualized results on the feature domain, compared with Chen *et al.* [21]. The mean feature maps are multiplied by 5 for better visualization. The lower intensity of our mean feature map is because we used a mean-pooling layer for the final feature fusion while Chen *et al.* used a max-pooling layer. The cow's dark eye is marked with red boxes.**

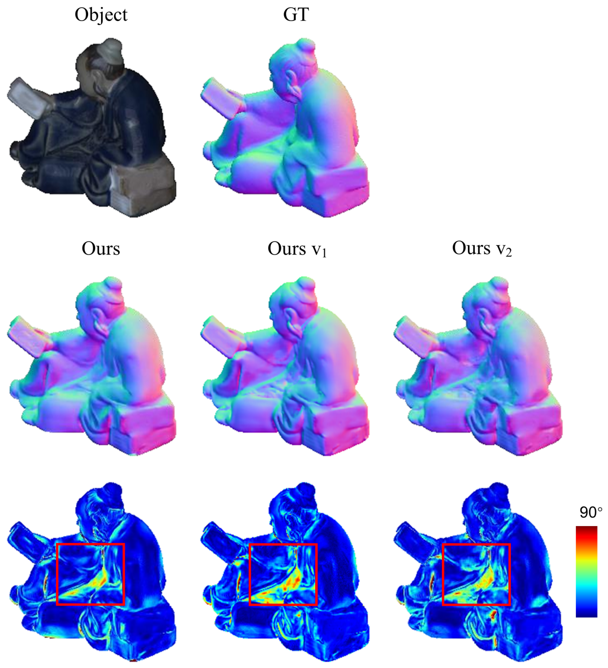

**Fig. 8. Visualized results of the ablation study. The object is the "Reading" from DiLiGenT benchmark [23]. The concave regions are marked with red boxes.**

only without pooling layers in modules. Ours v$_2$ represents the model without all feature fusion modules.

As shown in Table 2, our full method has the lowest errors for almost all of the samples and the lowest average errors on both real and synthetic data. The visualized results in Figure 8 also prove the improvement on concave regions. It is clearly shown that the feature fusion modules have an important effect on results, which proves the variation among multi-images can greatly improve the per-image feature extraction.

## 5. Conclusion

This paper proposed the inter-intra image feature fusion module for uncalibrated photometric stereo. With the proposed feature fusion module, the representations of inter-image variations are utilized to guide the per-image feature extraction, which makes the per-image local feature extraction more accurate and efficient for normal estimation.

The experimental results on the feature domain strongly demonstrate the effectiveness of our proposed feature fusion module. In addition, the quantitative and qualitative results show that our method performs significantly better on dark materials than the state-of-the-art method.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (Grant No.2019YFC1521200) and in part by the National Natural Science Foundation of China, No. 62172295.

## References

[1] R. J. Woodham, Photometric method for determining surface orientation from multiple images, Optical engineering 19 (1) (1980) 191139.

[2] F. Solomon, K. Ikeuchi, Extracting the shape and roughness of specular lobe objects using four light photometric stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (4) (1996) 449–454.

[3] S. Barsky, M. Petrou, The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1239–1252.

[4] M. Chandraker, S. Agarwal, D. Kriegman, Shadowcuts: Photometric stereo with shadows, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.

[5] Y. Mukaigawa, Y. Ishii, T. Shakunaga, Analysis of photometric factors based on photometric linearization, JOSA A 24 (10) (2007) 3326–3334.

[6] F. Verbiest, L. Van Gool, Photometric stereo with coherent outlier handling and confidence estimation, in: CVPR, IEEE, 2008, pp. 1–8.

[7] D. Miyazaki, K. Hara, K. Ikeuchi, Median photometric stereo as applied to the segonko tumulus and museum objects, International Journal of Computer Vision 86 (2) (2010) 229–242.

[8] C. Yu, Y. Seo, S. W. Lee, Photometric stereo from maximum feasible lambertian reflections, in: European Conference on Computer Vision, Springer, 2010, pp. 115–126.

[9] T.-P. Wu, C.-K. Tang, Photometric stereo via expectation maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (3) (2010) 546–560.

[10] A. S. Georghiades, Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo, in: ICCV, Vol. 3, IEEE Computer Society, 2003, pp. 816–816.

[11] H.-S. Chung, J. Jia, Efficient photometric stereo on glossy surfaces with wide specular lobes, in: CVPR, IEEE, 2008, pp. 1–8.

[12] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, Y. Ma, Robust photometric stereo via low-rank matrix completion and recovery, in: ACCV, Springer, 2010, pp. 703–717.

[13] S. Ikehata, D. Wipf, Y. Matsushita, K. Aizawa, Robust photometric stereo using sparse regression, in: CVPR, IEEE, 2012, pp. 318–325.

[14] N. G. Alldrin, S. P. Mallick, D. J. Kriegman, Resolving the generalized bas-relief ambiguity by entropy minimization, in: CVPR, IEEE, 2007, pp. 1–7.

[15] T. Papadhimitri, P. Favaro, A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima, International journal of computer vision 107 (2) (2014) 139–154.

[16] B. Shi, Y. Matsushita, Y. Wei, C. Xu, P. Tan, Self-calibrating photometric stereo, in: CVPR, IEEE, 2010, pp. 1118–1125.

[17] F. Lu, Y. Matsushita, I. Sato, T. Okabe, Y. Sato, Uncalibrated photometric stereo for unknown isotropic reflectances, in: CVPR, 2013, pp. 1490–1497.

[18] F. Lu, I. Sato, Y. Sato, Uncalibrated photometric stereo based on elevation angle recovery from brdf symmetry of isotropic materials, in: CVPR, 2015, pp. 168–176.

[19] G. Chen, K. Han, K.-Y. K. Wong, Ps-fcn: A flexible learning framework for photometric stereo, in: ECCV, 2018, pp. 3–18.

[20] G. Chen, K. Han, B. Shi, Y. Matsushita, K.-Y. K. Wong, Self-calibrating deep photometric stereo networks, in: CVPR, 2019, pp. 8739–8747.

[21] G. Chen, M. Waechter, B. Shi, K.-Y. K. Wong, Y. Matsushita, What is learned in deep uncalibrated photometric stereo?, in: ECCV, Springer, 2020, pp. 745–762.

[22] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, L. Van Gool, Uncalibrated neural inverse rendering for photometric stereo of general surfaces, in: CVPR, 2021, pp. 3804–3814.

[23] B. Shi, Z. Wu, Z. Mo, D. Duan, S.-K. Yeung, P. Tan, A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo, in: CVPR, 2016, pp. 3707–3716.

[24] S. Kumar, Y. Dai, H. Li, Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames, in: ICCV, 2017, pp. 4649–4657.

[25] J. L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: CVPR, 2016, pp. 4104–4113.

[26] Y. Furukawa, J. Ponce, Accurate, dense, and robust multiview stereopsis, IEEE transactions on pattern analysis and machine intelligence 32 (8) (2009) 1362–1376.

[27] S. Kumar, Y. Dai, H. Li, Superpixel soup: Monocular dense 3d reconstruction of a complex dynamic scene, IEEE transactions on pattern analysis and machine intelligence (2019).

[28] C. B. Choy, D. Xu, J. Gwak, K. Chen, S. Savarese, 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction, in: ECCV, Springer, 2016, pp. 628–644.

[29] O. Wiles, A. Zisserman, Silnet: Single-and multi-view reconstruction by learning from silhouettes, arXiv preprint arXiv:1711.07888 (2017).

[30] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, A. Bousseau, Flexible svbrdf capture with a multi-image deep network, in: Computer Graphics Forum, Vol. 38, Wiley Online Library, 2019, pp. 1–13.

[31] M. Aittala, F. Durand, Burst image deblurring using permutation invariant convolutional neural networks, in: ECCV, 2018, pp. 731–747.

[32] C. R. Qi, H. Su, K. Mo, L. J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: CVPR, 2017, pp. 652–660.

[33] F. Lu, X. Chen, I. Sato, Y. Sato, Symps: Brdf symmetry guided photometric stereo for shape and light source estimation, IEEE transactions on pattern analysis and machine intelligence 40 (1) (2017) 221–234.

[34] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[35] W. Matusik, H. Pfister, M. Brand, L. McMillan, A data-driven reflectance model, in: SIGGRAPH, 2003, pp. 759–769.