ODF-PSN: An Outlier-Inlier Dual-Region Decoupling Feature-Guided Network for Photometric Stereo

Zhihao Cai[1], Shiyu Qin[1], Yi Li[1], Lin Qi[1,*], and Junyu Dong[1,*]
[1] School of Computer Science and Technology, Ocean University of China, Qingdao, China
{caizhihao, qsy8264, Liyi_2025}@stu.ouc.edu.cn, {qilin, dongjunyu}@ouc.edu.cn

***Abstract*—Photometric stereo aims to recover surface normals from images captured under varying lighting conditions. Although deep learning methods have achieved promising results, existing approaches often struggle to accurately estimate surface normals in scenes where outlier and inlier regions are interwoven, due to entangled optical effects due to complex local structures and surface reflections. This results in a degradation of the accuracy of the reconstruction. To address this issue, we propose ODF-PSN, a dual-branch frame work that separately extracts features from outlier and inlier regions. Equipped with our Partitioned Feature Fusion Module, the network effectively integrates features from both regions. Furthermore, we introduce an anomaly-guided feature fusion module to enhance the representation of anomalous regions, thereby constructing a more robust global representation. Experimental results demonstrate that ODF-PSN achieves state-of-the-art performance on the DiLiGenT dataset.**

***Index Terms*—Photometric stereo, feature extraction, feature fusion.**

I. Introduction

Photometric Stereo [26] aims to reconstruct surface normals by analyzing intensity variations in images captured under different lighting conditions. Compared to other 3D reconstruction techniques [3] [5] [12], it offers finer surface details through physical modeling of light-shadow relationships.

Traditional photometric stereo methods [26] [20] [27] primarily rely on the ideal Lambertian assumption, modeling surfaces as perfectly diffuse. However, this rarely holds in practice; most real-world materials exhibit non-Lambertian behavior due to specular reflections, shadows, interreflections [17] [13], and complex surface microstructures. To address this, later approaches adopt more general reflectance models, particularly bidirectional reflectance distribution functions (BRDFs) [21], to capture diverse lighting–viewing interactions. Despite their expressiveness, BRDF-based methods often require strong material priors, assume calibrated lighting, and remain sensitive to noise and sparse observations, leading to unstable optimization and poor generalization in real-world settings.

To handle non-Lambertian surfaces, deep learning-based approaches [19] [6] [2] have been widely adopted and have significantly advanced the state of the art in photometric stereo. However, as shown in Fig. 1, existing methods still struggle with regions where the outlier and inlier region coexist, areas that often contain critical geometric and reflectance information. These challenges arise due to overexposure, underexposure, or strong non-linear responses when applying uniform feature extraction strategies.In this paper, we define the regions with highlights and shadows as outlier regions, while other areas are considered as inlier regions.

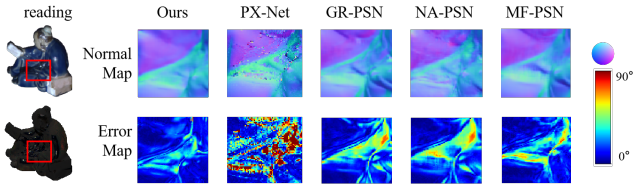

Fig. 1: An example of the errors in complex-structured regions. We visualized the error plot of the area where outlier areas and inlier areas were mixed. The results compare our method with the PX-Net [15] and GR-PSN [10] and NA-PSN [11] and MF-PSN [14].

To address this issue, we propose ODF-PSN. Our main contributions are as follows:

- We present a dual-region decoupling framework that explicitly separates outlier and inlier regions and processes them through partitioned feature extraction module, enabling tailored feature extraction for their distinct reflectance properties.
- We design a partitioned attention-based feature fusion module that adaptively combines features from both branches by dynamically weighting discriminative regions according to input content, yielding more robust and context-aware representations.
- We introduce an outlier-guided fusion module that effectively integrates fine-grained local details from outlier regions with global contextual semantics, facilitating unified optimization across multi-level features.

Extensive experiments demonstrate that ODF-PSN achieves state-of-the-art performance on the DiLiGenT benchmark with a mean angular error (MAE) of 6.18° and exhibits superior robustness under challenging photometric conditions, including overexposure, underexposure, and strong nonlinear sensor responses. Our code and models can be found at

*Corresponding authors: Lin Qi and Junyu Dong.

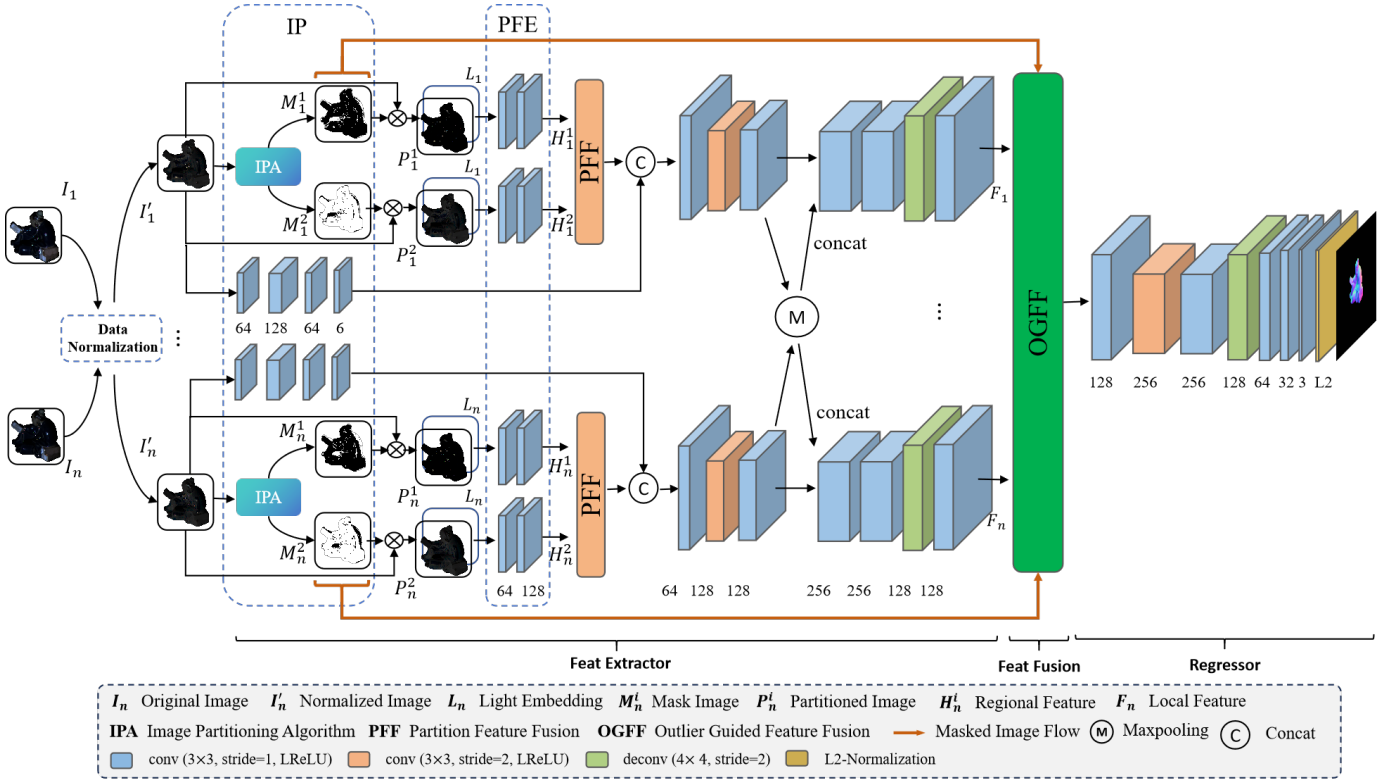

Fig. 2: ODF-PSN consists of a shared weight feature extraction subnet, a fusion layer and a normal regression network.

https://github.com/dy6666/ODF-PSN.git.

II. Method

A. ODF-PSN

The network architecture of ODF-PSN is illustrated in Fig. 2, comprises a weight-shared feature extractor, a feature fusion module, and a normal regression module. From an RGB input $X \in \mathbb{R}^{3\times H\times W}$, the image is partitioned and concatenated with the light direction vector to form $F \in \mathbb{R}^{6\times H\times W}$. This tensor is processed by the Partitioned Feature Extraction (PFE) and Fusion (PFF) modules to yield a unified representation, which is then combined with the original features and refined. The Outlier-Guided Feature Fusion (OGFF) module enhances anomalous regions while integrating global context, and the regression module finally outputs the surface normal map $N \in \mathbb{R}^{3\times H\times W}$.

B. Preliminaries

1) Data Normalization: To mitigate the adverse effects of spatially varying BRDFs (SVBRDFs) on feature learning, we normalize the input images following the image formation model [21] [4]. Given multiple observations per pixel across different lighting conditions, this simple normalization effectively suppresses albedo variations and reduces ambiguity between reflectance and geometry, thus reducing the negative impact of SVBRDFs on normal estimation.

2) Image Partitioning: We propose Image Partitioning (IP) which splits normalized input images into inlier and outlier regions using our Image Partitioning Algorithm (IPA). IPA employs two fixed thresholds, a high threshold of 0.25 and a low threshold of 0.02 chosen empirically to best separate outliers from inliers. A pixel is labeled a highlight if any RGB channel exceeds 0.25 or a shadow only if all channels fall below 0.02. Highlights and shadows together form the outlier region, while the remaining pixels constitute the inlier region. For efficient downstream processing, we generate binary masks for each image, where a value of 1 denotes inclusion in the target region and 0 indicates exclusion. This yields two complementary masks, the outlier mask and the inlier mask. The efficacy of our partitioning strategy is visually demonstrated in Fig. 3.

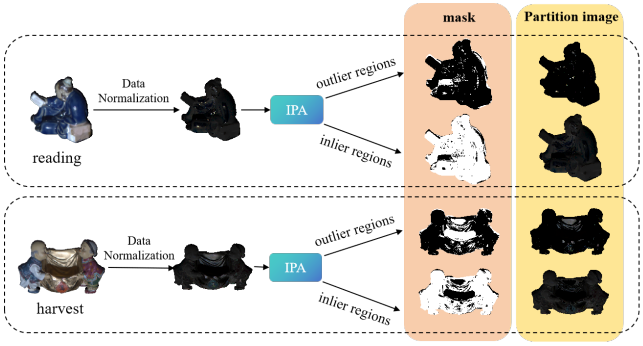

Fig. 3: The result of image partitioning, “mask” represents the mask after partitioning, and “partitioning image” represents the result of image partitioning.

C. Partitioned Feature Extraction Module(PFE)

To address the issue that using a unified feature extractor for both outlier regions and inlier regions would compromise the reconstruction accuracy, we employ separate feature extractors to process different regions independently. As shown in Fig.2, in the PFE component, we utilize two convolutional layers that preserve the original image resolution (H, W) to extract features from each region, effectively avoiding mutual interference between using a shared feature extractor for mixed regions. Through this process, we obtain two distinct feature maps, denoted as H_i^1 and H_i^2, representing the feature of outlier areas and inlier areas, respectively.

D. Partitioned Feature Fusion Module(PFF)

For the region-partitioned features generated by PFE, simple concatenation or addition is insufficient for effective fusion. To address this limitation, we design a partitioned feature fusion module that leverages attention to dynamically assign higher weights to more discriminative regions based on the input content, enabling adaptive integration of features from both branches.

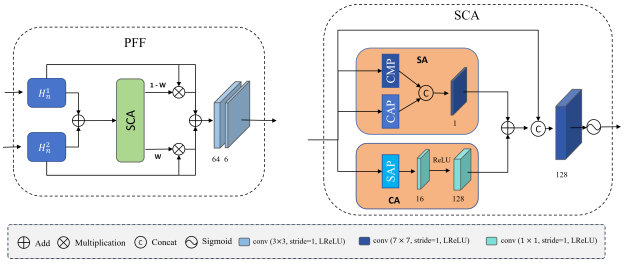

Fig. 4: The Architecture of Partitioned Feature Fusion Module

Given that simple concatenation or summation fails to effectively fuse features, we propose a Partitioned Feature Fusion Module (PFF), whose structure is shown in Fig. 4. After extracting regional features, the outlier region features and inlier region features are first summed to obtain initial features. These initial features are then passed through a Spatial-Channel Attention (SCA) module to generate feature weights for both regions. The weighted features are multiplied by their respective weights and summed to produce first-level fused features. Finally, the first-level fused features are added to the initial features to obtain the final fused features, as defined below:

$$\begin{aligned} \mathbf{S}_i &= \mathbf{H}_i^1 + \mathbf{H}_i^2 \\ w &= \mathrm{SCA}(\mathbf{S}_i) \\ \mathbf{F}_i^t &= \mathbf{H}_i^1 \cdot (1-w) + \mathbf{H}_i^2 \cdot w + \mathbf{S}_i \\ \mathbf{T}_i &= \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{3\times3}(\mathbf{F}_i^t)) \end{aligned} \tag{1}$$

where H_i^1 and H_i^2 denote the features of the outlier region and the inlier region in the i-th pair of features, respectively. SCA denotes the attention weighter. w represents the weights at different positions. ReLU denotes relu function. Conv denotes the convolutional layer.

The SCA module integrates both channel-wise and spatial attention mechanisms. Specifically, channel attention generates per-channel weights W_c to recalibrate feature channels, while spatial attention computes a spatial importance map W_s to highlight critical spatial regions. These attention weights are combined via element-wise addition and passed through a sigmoid activation to produce the final adaptive weights for outlier and inlier region features, as follows:

$$\begin{aligned} \mathrm{W}_s &= \mathrm{Conv}_{7\times7}(\mathrm{Concat}(\mathrm{CAP}(\mathbf{S}_i), \mathrm{CMP}(\mathbf{S}_i)) \\ \mathrm{W}_c &= \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}(\mathrm{SAP}(\mathrm{S}_i)))) \\ \mathrm{W} &= \sigma\big(\mathrm{Conv}_{7\times7}(\mathrm{Concat}(\mathrm{W}_c + \mathrm{W}_s, \mathrm{S}_i))\big) \end{aligned} \tag{2}$$

where S_i represents the summation of features of the outlier and inlier regions. CMP represents the global Max pooling operation in the channel dimension. CAP denotes the average pooling operation in the channel dimension. SAP stands for the average pooling operation across the spatial dimensions. The sigmoid function is denoted by σ, and Concat refers to the operation that concatenates features along the channel dimension. ReLU denotes the relu function.

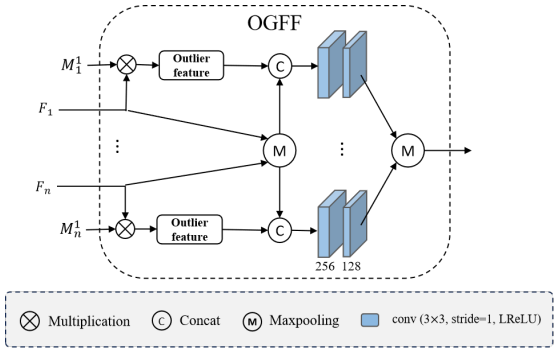

Fig. 5: The Architecture of outlier-guided feature fusion

E. Outlier-Guided Feature Fusion Module(OGFF)

Partitioned feature extraction and fusion effectively decouple outlier and inlier features, preventing mutual interference. However, because outlier regions often exhibit complex photometric effects such as highlights, shadows, and interreflections, the extracted features from these regions are typically of limited quality.

To improve their representation, we propose an outlier-guided feature fusion module, as shown in Fig. 5. First, the global features are fused through max-pooling to generate a foundational global feature representation f. Then, the global inlier features are enhanced by using local outlier region features f_i^m. Specifically, the global base features are concatenated with local outlier region features and processed through convolutional layers to produce enhanced local features. These features are then max-pooled to derive the final global feature G. This design maintains the consistency of global features while strengthening the representation of key region details. The formal definition is as follows:

$$
\begin{aligned}
\mathbf{f} &= \text{Maxpool}(\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_n) \\
\mathbf{f}_i^m &= \mathbf{F}_i \cdot \mathbf{M}_i^1 \\
\mathbf{L}_i &= \text{Conv}_{3\times3}(\text{Conv}_{3\times3}(\text{Concat}(\mathbf{f}, \mathbf{f}_i^m))) \\
\mathbf{G} &= \text{Maxpool}(\mathbf{L}_1, \mathbf{L}_2, \cdots \mathbf{L}_n)
\end{aligned} \tag{3}
$$

where F_i represents the local feature extracted from the i-th image. M_i^1 are the masks that identify outlier regions, respectively. The Maxpool operation is used to fuse all local features by performing max-pooling across these regions.

F. Loss Function

The network is trained in a fully supervised manner by optimizing parameters based on the angular discrepancy between predicted surface normals and ground truth normals.

$$
Loss = \frac{1}{HW} \sum_{i,j}^{HW} (1 - cos(\tilde{N}_{i,j} \cdot N_{i,j})) \tag{4}
$$

where H, W represents the image resolution, $N_{i,j}$ denotes the ground truth normal vector at pixel (i,j), and $\tilde{N}_{i,j}$ indicates the predicted normal vector at pixel (i,j). The predicted normal vector and ground truth normal vector are combined through a dot product to reflect their angular relationship. Values closer to 1 indicate smaller angular differences between normals, while values farther from 1 suggest larger discrepancies.

III. Experiments

A. Experimental setup

Dataset: We use two types of datasets: training and testing. For training, we mainly employ the Blobby [7] and Sculpture [25] datasets, both synthesized from 3D models using the MERL material database [16], which contains 100 real-world materials. Each object is rendered with two randomly selected materials, yielding samples of 64 images under distinct lighting directions. The dataset are split into training and validation sets at a 99:1 ratio. For evaluation, we employ two real-world benchmarks: DiLiGenT [20] and DiLiGenT-Π [22]. DiLiGenT includes 10 objects with complex surfaces, each captured under 96 lighting conditions. DiLiGenT-Π contains 30 objects with even richer surface details, each imaged under 100 lighting conditions.

Training details: The experiments were conducted on a single NVIDIA GeForce RTX 3090 24GB GPU using PyTorch. The model was trained for 30 epochs using the Adam optimizer (β_1=0.9, β_2=0.999) with an initial learning rate of 0.0012, which was halved every 5 epochs. The total training time was approximately 14 hours.

Evaluation metrics: The primary evaluation metric is the Mean Angular Error (MAE), measured in degrees. MAE quantifies the average angular deviation between the predicted and ground-truth surface normals across all pixels in the image. Specifically, it is calculated as follows:

$$
MAE = \frac{1}{HW} \sum_{i,j}^{HW} (arccos(\tilde{N}_{i,j} \cdot N_{i,j})) \tag{5}
$$

where H, W denotes the image resolution. n_p represents the surface normal vector of ground truth, while $\tilde{n}_p$ indicates the predicted surface normal vector. Their angular difference is computed via the arccos function of their dot product.Additionally, we introduced two evaluation metrics, err^{15° and err^{30° [9]. err^{15° denotes the proportion of surface normals with mean angular errors less than or equal to 15 degrees, while err^{30° represents the proportion with mean angular errors less than or equal to 30 degrees.

B. Benchmark Comparisons

Comparative Evaluation on DiLiGenT Benchmark: To comprehensively evaluate our method and demonstrate its superiority, we conducted extensive experiments on the DiLiGenT benchmark dataset [20] comparing it with ten state-of-the-art photometric stereo methods. As shown in Table I, our method achieves a mean angular error of 6.18 degrees outperforming all competing approaches. It attains the best results on challenging objects such as Buddha Harvest and Reading showing strong robustness under extreme lighting. The qualitative results in Fig. 6 further highlight our model's superior ability to reconstruct complex surfaces especially in mixed regions containing both inlier and outlier orientations. These include the earlobe of "Buddha" the pocket area of "Harvest" and the abdominal folds of "Reading" where transitional lighting self-occlusion or fine geometric structures cause complex appearance variations. The performance gain is confirmed by significantly lower angular errors in these areas. Moreover our approach excels on hybrid surfaces and transitional zones near specular highlights demonstrating strong geometric coherence under extreme illumination. These results validate that our feature extraction and fusion strategy effectively captures fine-grained details in structurally complex and photometrically challenging scenarios.

TABLE I: Comparisons on DiLiGenT-main benchmark [20]. Red indicates the best result and blue indicates the second best result.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
PS-FCN(N.) [1]	2.67	7.72	7.53	4.76	6.72	7.70	12.39	6.17	7.15	10.92	7.39
MF-PSN [14]	2.07	5.83	6.88	5.00	5.90	7.46	13.38	7.20	6.81	12.20	7.27
CNN-PS [6]	2.12	4.38	8.07	4.38	7.92	7.42	14.08	5.37	6.38	12.12	7.20
PSMF-PSN [28]	2.54	5.99	7.21	5.09	5.52	7.75	11.40	6.91	6.11	10.01	6.85
CFR-PSN [24]	2.08	5.65	7.28	4.19	7.93	7.34	11.66	5.27	6.70	9.98	6.81
SR-PSN [8]	2.23	5.24	6.75	4.63	6.12	7.07	12.61	5.88	6.44	10.35	6.73
NA-PSN [11]	2.93	4.65	7.12	4.65	5.99	7.49	12.28	5.96	6.42	9.93	6.72
GR-PSN [10]	2.22	5.61	6.73	4.33	6.17	6.78	12.03	5.54	6.42	9.65	6.55
IGA-PSN [23]	2.20	4.83	7.08	4.63	5.92	7.25	11.01	5.85	6.41	9.37	6.46
PX-Net [15]	2.03	4.13	7.61	4.39	4.69	6.90	13.10	5.08	5.10	10.26	6.33
MSF-Net [18]	2.27	4.60	6.84	4.20	6.08	7.40	10.80	5.08	5.85	9.15	6.23
ODF-PSN(Ours)	2.13	4.22	6.60	4.41	6.30	6.62	11.04	5.58	6.13	8.76	6.18

Comparative Evaluation on DiLiGenT-Π Benchmark: Reconstruction of fine shape details remains a key challenge in photometric stereo, especially for real-world objects such as reliefs, badges, and coins. On the DiLiGenT-Π dataset [22], ODF-PSN achieves a mean angular error (MAE) of 9.4°, with the most significant improvement observed on "rough"

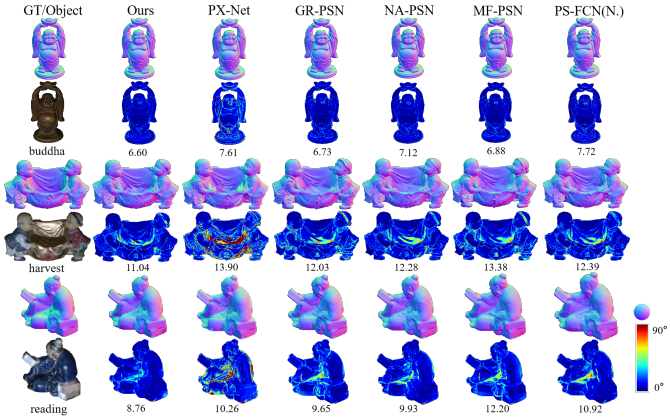

Fig. 6: Quantitative analysis on the "buhhda","harvest", and "reading" objects from the DiLiGenT dataset, using 96 images under different lighting directions as input. The numbers below the error plots represent the Mean Absolute Error (MAE). We compared our method with several state-of-the-art methods, including PX-Net [15], GR-Net [10], NA-PSN [11], MF-PSN [14], and PS-FCN(N.) [1].

TABLE II: Benchmark results on our real-world dataset DiLiGenT-Π [22]. We calculate the average MAE of the object in each material.

material	L2 [26]	PX-Net [15]	PS-FCN [2]	GPS-Net [29]	CNN-PS [6]	NA-PSN [11]	Ours
Metallic	7.2	9.2	5.2	5.0	4.9	4.9	5.3
Specular	8.5	8.4	8.5	8.9	7.4	7.6	7.8
Translucent	17.0	16.0	17.3	17.6	16.8	17.0	17.2
Rough	14.6	13.7	14.5	13.8	13.5	13.3	13.2
AVG.	10.5	10.8	9.8	10.1	9.2	9.2	9.4

materials, where it outperforms all competing methods by a clear margin. This advantage arises because rough surfaces inherently exhibit dense intermixing of outlier regions, such as localized highlights and cast shadows, with inlier regions, creating highly ambiguous reflectance patterns that challenge conventional approaches. By explicitly decoupling these regions and applying dedicated feature extraction and fusion strategies, our method effectively disentangles their distinct photometric behaviors, enabling accurate recovery of sharp geometric transitions and subtle depth variations. The consistent performance across diverse object types, including textured reliefs and engraved patterns, further validates that ODF-PSN's outlier-inlier aware design is well suited for real-world photometric stereo under complex non-Lambertian reflectance.

C. Ablation study

In our ablation studies, we adopt three primary evaluation metrics: Mean Angular Error (MAE), err^{15°, and err^{30°, to assess model performance. Lower MAE values indicate superior accuracy, whereas higher err^{15° and err^{30° scores are more favorable. As presented in Table III, the experiments evaluate three core architectural components across six different configurations in a systematic manner. Notably, the integration of the Partitioned Feature Fusion (PFF) module depends on the presence of the Partitioned Feature Extraction (PFE) module, which introduces certain limitations on feasible experimental combinations. The six experimental configurations are defined as follows: ID0 serves as the baseline configuration without any additional components; ID1 introduces the partitioned feature extraction module based on ID0; ID2 incorporates the partitioned feature fusion module upon ID1; ID3 implements the outlier-guided feature fusion module starting from ID0; ID4 adds the outlier-Guided feature fusion module to ID1; ID5 represents the complete model incorporating all components.

TABLE III: Ablation results on DiLiGenT benchmark [20], in terms of MAE, the proportion of pixels with MAE less than 15° and 30°. PFE: Partitioned Feature Extraction; PFF: Partitioned Feature Fusion; OGFF: Outlier-Guided Feature Fusion. (✓: used, - : not used).

ID	PFE	PFF	OGFF	MAE ↓	err^{15° ↑	err^{30° ↑
0	–	–	–	23.39	25.6%	62.6%
1	✓	–	–	6.48	92.1%	98.0%
2	✓	✓	–	6.46	92.2%	98.1%
3	–	–	✓	26.57	38.3%	67.1%
4	✓	–	✓	6.41	92.4%	98.1%
5	✓	✓	✓	6.18	92.7%	98.1%

1) Effectiveness of partition feature extraction: The experimental comparison between ID0 and ID1, as well as ID3 and ID4, demonstrates that the introduction of PFE module leads to significant improvements in MAE, err^{15° and err^{30°. As shown in the results of ID0 and ID1, using only non-partitioned images for initial feature extraction and compressing the features into 6 channels as guidance results in limited feature dimensionality and disorganized representation, leading to poor surface normal estimation. In contrast, with the PFE module, the feature representational capacity is substantially enhanced, yielding a reduction 16.91° in MAE. Further analysis of ID3 and ID4 reveals that without PFE to establish clear regional separation, the OGFF module alone struggles to effectively integrate features. However, when OGFF is used in conjunction with PFE, it can fully leverage structurally guided features, resulting in a reduction 20.16° in MAE. These results comprehensively validate the effectiveness of the PFE module.

2) Effectiveness of partition feature fusion: The experimental comparison between ID1 and ID2, as well as ID4 and ID5, shows that the introduction PFF module reduces MAE by 0.02° and 0.23°, with slight improvements in err^{15° and err^{30°. These findings indicate an overall reduction in angular errors, with the error distribution becoming more concentrated within the 15° range, and large errors originally exceeding 30° gradually shifting toward smaller error bins. In particular, merely performing partitioned feature extraction without effective fusion fails to fully exploit the structured information introduced by region partitioning. These results effectively validate the efficacy of the proposed feature fusion component and underscore its critical role in enabling coherent and context-aware feature integration within our network architecture.

3) Effectiveness of outlier-guided feature fusion: The comparison between ID0 and ID3 shows that, when OGFF is used alone, the MAE increases, while both err^{15° and err^{30° improve—indicating that the module effectively enhances normal estimation for the majority of pixels. The rise in MAE is likely caused by a small number of outliers with large angular errors, possibly due to over-enhancement of outlier regions in the absence of structured guidance. In contrast, comparisons between ID1 and ID4, as well as ID2 and ID5, show that applying OGFF on top of PFE reduces MAE by 0.07° and 0.28°, respectively, while also yielding modest improvements in both err^{15° and err^{30°. This improvement comes from the fact that PFE provides clear region separation, enabling OGFF to effectively focus on critical anomalous regions and perform targeted feature enhancement. These results confirm that OGFF acts as a refinement module whose effectiveness is contingent on structured input features, thus validating its design and role in the overall framework.

IV. Conclusions

In this paper, We propose an Outlier-inlier Dual-Region Decoupled Feature-Guided Network for photometric stereo. Our method first separates input images into outlier and inlier regions and extracts their features independently. A partitioned feature fusion module integrates these region-specific features, followed by a dual-stage fusion module that selectively enhances global inlier features using local outlier cues and performs secondary fusion. Surface normals are then predicted via a regression network. Extensive ablation studies and benchmark comparisons validate the effectiveness and superiority of our approach.

Acknowledgment

This study was supported by the Major Program of Shandong Provincial Natural Science Foundation (Grant No. ZR2024ZD04) and the Key R&D Program of Shandong Province, China (Grant No. 2024ZLGX06).

References

[1] Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Deep photometric stereo for non-lambertian surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 129–142 (2020)

[2] Chen, G., Han, K., Wong, K.Y.K.: PS-FCN: A flexible learning framework for photometric stereo. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11205, pp. 3–19. Springer, Cham, Switzerland (2018)

[3] Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(12), 4338–4364 (2020)

[4] Holroyd, M., Lawrence, J., Humphreys, G., Zickler, T.: A photometric approach for estimating normals and tangents. ACM Transactions on Graphics **27**(5), 1–9 (Dec 2008)

[5] Huang, B., Zhao, J., Liu, J.: A survey of simultaneous localization and mapping. arXiv preprint arXiv:1909.05214 **1**(2.3) (2019)

[6] Ikehata, S.: CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11205, pp. 20–37. Springer, Cham, Switzerland (2018)

[7] Johnson, M.K., Adelson, E.H.: Shape estimation in natural illumination. In: CVPR 2011 (Jun 2011)

[8] Ju, Y., Jian, M., Wang, C., Zhang, C., Dong, J., Lam, K.M.: Estimating high-resolution surface normals via low-resolution photometric stereo images. IEEE Transactions on Circuits and Systems for Video Technology **34**(4), 2512–2524 (2024)

[9] Ju, Y., Lam, K.M., Chen, Y., Qi, L., Dong, J.: Pay attention to devils: A photometric stereo network for better details. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (Jul 2020)

[10] Ju, Y., Shi, B., Chen, Y., Zhou, H., Dong, J., Lam, K.M.: Gr-psn: Learning to estimate surface normal and reconstruct photometric stereo images. IEEE Transactions on Visualization and Computer Graphics pp. 1–16 (2023)

[11] Ju, Y., Shi, B., Jian, M., Qi, L., Dong, J., Lam, K.M.: Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention. International Journal of Computer Vision **130**(12), 3014–3034 (2022)

[12] Li, Y., Dong, J., Dong, Z., Yang, C., An, Z., Xu, Y.: Srkd: Towards efficient 3d point cloud segmentation via structure-and relation-aware knowledge distillation. arXiv preprint arXiv:2506.17290 (2025)

[13] Li, Y., Meng, S., Yang, C., Feng, W., Liu, J., An, Z., Wang, Y., Tian, Y.: A comprehensive survey of interaction techniques in 3d scene generation. Authorea Preprints (2026)

[14] Liu, Y., Ju, Y., Jian, M., Gao, F., Rao, Y., Hu, Y., Dong, J.: A deep-shallow and global–local multi-feature fusion network for photometric stereo. Image and Vision Computing p. 104368 (Feb 2022)

[15] Logothetis, F., Budvytis, I., Mecca, R., Cipolla, R.: Px-net: Simple, efficient pixel-wise training of photometric stereo networks (Aug 2020)

[16] Matusik, W., Pfister, H., Brand, M., McMillan, L.: A data-driven reflectance model. In: ACM SIGGRAPH 2003 Papers (Jul 2003)

[17] Nayar, S.K., Ikeuchi, K., Kanade, T.: Shape from interreflections. International Journal of Computer Vision p. 173–195 (Jul 1991)

[18] Qin, S., Cai, Z., Wang, K., Qi, L., Dong, J.: Msf-net: Multi-stage feature extraction and fusion for robust photometric stereo. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 425–438. Springer (2025)

[19] Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) (Oct 2017)

[20] Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S.K., Tan, P.: 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)

[21] Shi, B., Tan, P., Matsushita, Y., Ikeuchi, K.: Bi-polynomial modeling of low-frequency reflectances. IEEE Transactions on Pattern Analysis and Machine Intelligence p. 1078–1091 (Jun 2014)

[22] Wang, F., Ren, J., Guo, H., Ren, M., Shi, B.: Diligent-pi: A photometric stereo benchmark dataset with controlled shape and material variation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9477–9487 (October 2023)

[23] Wang, K., Qi, L., Qin, S., Luo, K., Ju, Y., Li, X., Dong, J.: Image gradient-aided photometric stereo network. In: Hadfi, R., Anthony, P., Sharma, A., Ito, T., Bai, Q. (eds.) PRICAI 2024: Trends in Artificial Intelligence. pp. 284–296. Springer Nature Singapore, Singapore (2025)

[24] Wang, Z., Gao, Y., Li, X., Gu, P., Zhao, B., Li, X.: A coarse-to-fine reconstruction framework for non-lambertian photometric stereo. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2024)

[25] Wiles, O., Zisserman, A.: Silnet: Single- and multi-view reconstruction by learning from silhouettes. In: Procedings of the British Machine Vision Conference 2017 (Jan 2017)

[26] Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical Engineering (Feb 1980)

[27] Wu, L., Ganesh, A., Shi, B., Matsushita, Y., Wang, Y., Ma, Y.: Robust photometric stereo via low-rank matrix completion and recovery. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) Computer Vision – ACCV 2010. pp. 703–717. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)

[28] Yang, Y., Liu, J., Ni, Y., Li, C., Wang, Z.: Accurate normal measurement of non-lambertian complex surface based on photometric stereo. IEEE Transactions on Instrumentation and Measurement **72**, 1–11 (2023)

[29] Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B.: Gps-net: Graph-based photometric stereo network (Jan 2020)