EFFICIENT FEATURE FUSION FOR LEARNING-BASED PHOTOMETRIC STEREO

Yakun Ju[†] *Kin-Man Lam*[†] *Jun Xiao*[†] *Cong Zhang*[†] *Cuixin Yang*[†] *Junyu Dong*[‡]

[†] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University
[‡] School of Computer Science and Technology, Ocean University of China

ABSTRACT

How to handle an arbitrary number for input images is a fundamental problem of learning-based photometric stereo methods. Existing approaches adopt max-pooling or observation map to fuse an arbitrary number of extracted features. However, these methods discard a large amount of the features from the input images, impacting the utilization and accuracy, or ignore the constraints from the intra-image spatial domain. In this paper, we explore how to efficiently fuse features from a variable number of input images. First, we propose a bilateral extraction module, which categorizes features into positive and negative, to maximally keep the useful feature in the fusion stage. Second, we adopt a top-k pooling to both the bilateral information, which selects the k maximum response value from all features. These two modules proposed are "plug-and-play" and can be used in different fusion tasks. We further propose a hierarchical photometric stereo network, namely HPS-Net, to handle bilateral extraction and top-k pooling for multiscale features. Experiments in the widely used benchmark illustrate the improvement of our proposed framework in the conventional max-pooling method and the proposed HPS-Net outperforms existing learning-based photometric stereo methods.

***Index Terms*—** Photometric stereo, deep neural network, feature fusion

1. INTRODUCTION

Three-dimensional (3D) shape is a fundamental problem in computer vision since it will facilitate the understanding of two-dimensional (2D) images. Photometric stereo is a single-view 3D shape recovery technique and prevails in recovering high-resolution dense surface normals. Although the earliest method solved Lambertian surfaces [1], most real-world objects are non-Lambertian. Conventional works manily deal with these general surfaces by approximating the bidirectional reflectance distribution function (BRDF) [2] or rejecting non-Lambertian outliers [3]. However, these methods are accurate for limited categories of materials and suffer from unstable non-convex optimization [4].

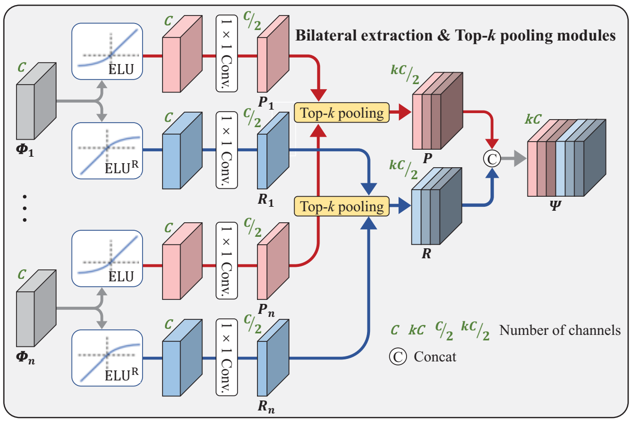

Fig. 1. Architecture of the proposed bilateral extraction module and top-k pooling module.

Meanwhile, deep learning techniques [5, 6] have shown powerful fitting ability in photometric stereo networks, which focus on the reconstruction under non-Lambertian surfaces [7, 8]. However, the effectiveness of feature aggregation in photometric stereo is rarely discussed. It is very difficult for the photometric stereo networks to have the ability of handling an order-agnostic and variable number of input images required in real photometric stereo task [9]. In fact, this problem is equivalent to how to fuse an unfixed number of features in the networks. It is known that convolutional neural networks (CNNs) are incapable of handling a variable number of inputs during training and testing. Previous learning-based methods only adopt two limited strategies for variable fusion: the observation map approach and the feature max-pooling approach. The observation map approach focuses on single-pixel features, aggregating each pixel of all input images under different projected light directions into a fixed-size feature map [7]. However, this single-pixel-based method isolates the spatial constraint within the images, and cannot be applied to the uncalibrated photometric stereo task.

On the other hand, feature max-pooling pays more attention to the features of each image [8]. However, the tricky

The work was supported by the Key-Area Research and Development Program of Guangdong Province (2020B090928001), the Project of Strategic Importance Fund from The Hong Kong Polytechnic University (No. ZE1X), and the National Key Scientific Instrument and Equipment Development Projects of China (41927805).

problem is that this approach only retains the maximum response at each position, thus discarding a large amount of information in the inputs. Therefore, how to better fuse the order-agnostic and variable number of features is a challenging problem, which significantly affects the performance of existing deep learning-based photometric stereo networks.

To achieve this goal, we propose two simple but effective bilateral extraction and top-k pooling modules. Fig. 1 show these two modules. Given n arbitrary features, $\boldsymbol{\Phi}_1$, $\boldsymbol{\Phi}_2$, $\cdots$, $\boldsymbol{\Phi}_n$, learned from a extractor, the proposed bilateral extraction module first outputs positive and negative information of each feature, by rotating 180° clockwise (equivalent to multiplying the feature by -1), inspired by [10]. Afterwards, we adopt a top-k pooling, to select the k maximum response values from the n positive features $\boldsymbol{P_1}$, $\boldsymbol{P_2}$, $\cdots$, $\boldsymbol{P_n}$, and the n rotated negative features $\boldsymbol{R_1}$, $\boldsymbol{R_2}$, $\cdots$, $\boldsymbol{R_n}$.

Inspired by HR-Net [11], we also develop a hierarchical photometric stereo network (HPS-Net), which adopts a multiscale structure in the feature extractor, as shown in Fig. 2. Our proposed HPS-Net has two advantages: (1) We employ a parallel network structure to extract features at three scales to avoid learning feature maps from low to high resolutions. Therefore, HPS-Net can generate features of high semantics and with high resolution details for the estimation of surface normals. (2) The feature extractor can provide the bilateral extraction and the top-k pooling modules with multiscale features. This can further retain the useful information for the stage of normal regression. Extensive experiments show that our proposed modules can improve the performance of existing max-pooling-based networks, and the proposed HPS-Net achieves superior performance over state-of-the-art methods.

2. METHOD

2.1. Bilateral Extraction

Conventional activation functions, such as ReLU, LeakyReLU, and ELU, simply truncate or attenuate the negative parts of features. These functions can lead to undesired information loss or distortion. In fact, retaining the negative parts in an inhibited state may be conducive to some high-level tasks, such as segmentation and detection. However, in the regression tasks (*e.g.*, photometric stereo, monocular depth estimation, *etc.*), those non-activated parts may have features needed for regression. Although some activation functions, such as Sigmoid, Tanh, have the ability to respond globally, the gradient in the saturation regions is very small, which easily causes the problem of vanishing gradient.

To overcome the aforementioned limitations, we adopt bilateral extraction, inspired by [10], to maximize the sue of the negative parts of features, while retaining non-linearity. As can be seen in Fig. 1, the feature $\boldsymbol{\Phi}_i$ where $i = 1, 2, \cdots, n$, will be activated by an original and a 180° clockwise rotated activation function (bilateral activation), forming the positive feature $\boldsymbol{P_i}$ and the rotated negative feature $\boldsymbol{R_i}$.

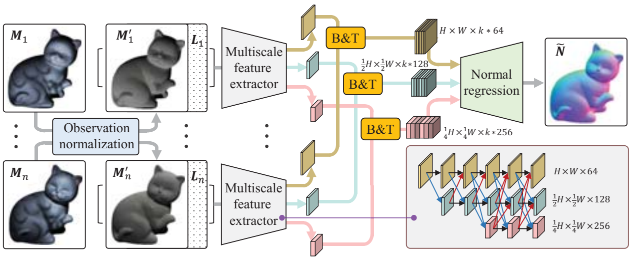
M_1
M'_1
L_1
Multiscale feature extractor
B&T
$H \times W \times k * 64$
$\frac{1}{2}H \times \frac{1}{2}W \times k * 128$
B&T
Normal regression
$\tilde{N}$
B&T
$\frac{1}{4}H \times \frac{1}{4}W \times k * 256$
Observation normalization
M_n
M'_n
L_n
Multiscale feature extractor
$H \times W \times 64$
$\frac{1}{2}H \times \frac{1}{2}W \times 128$
$\frac{1}{4}H \times \frac{1}{4}W \times 256$

Fig. 2. The architecture of the proposed HPS-Net. B&T represents the proposed bilateral extraction and top-k pooling.

Another advantage of bilateral activation is shown by the subsequent k pooling operation (see Section 2.2). The positive features and rotated negative features can provide one more group of pooling, which definitely preserves more useful features for normal regression.

2.2. Top-k Pooling

To enable the module to aggregate an arbitrary number of input features, we utilize a top-k pooling operation after bilateral activation, which retains k maximum values at the same position from all the inputs. This operation can extract and retain k times more information than the original max-pooling, which is beneficial to the reconstruction of surface normals. In fact, max-pooling can extract the most salient information from all the features, thereby keeping strong clues in regions of high intensities or specular highlights [12]. Therefore, we choose the method of ranking the maximum, rather than fusing the average value (average-pooling) or ranking the minimum values (min-pooling).

2.3. HPS-Net

We develop a hierarchical photometric stereo network, namely HPS-Net, using the above mentioned modules. As shown in Fig. 2, HPS-Net can be divided into four parts: observation normalization, multiscale feature extractor, multiscale bilateral extraction & top-k pooling, and normal regression.

Observation normalization: A CNN-based photometric stereo framework, which handles the patch-level inputs and is trained with homogeneous BRDF, may fail to estimate the input with spatially varying colors [12]. Therefore, we employ an observation normalization method [13, 12] to remove the impact of spatially varying BRDFs, that are common in real-world objects. The operation is to normalize each observation by all the n observations, as follows:

$$m'_i = \frac{m_i}{\sqrt{m_1^2 + m_2^2 + \cdots + m_n^2}}, \ i \in \{1, 2, \cdots, n\}, \quad (1)$$

where m_i and m'_i represent a pixel value in the original observation $\boldsymbol{M_i}$ and the normalized observation $\boldsymbol{M'_i}$, respectively.

It can be seen that, under the assumption of Lambertian reflectance [1], the ρ in $m_i = \rho\, \boldsymbol{n}^{\top}\boldsymbol{l}$ can be removed, where $\boldsymbol{n}$ and $\boldsymbol{l}$ represent the surface normal and incident light.

Multiscale feature extractor: We first repeat the xyz 3-dimensional light direction vector $\boldsymbol{l_i}$ to form an expanded $\boldsymbol{L_i}$ $\in \mathbb{R}^{H\times W\times 3}$ having the same spatial size as the observations, and then concatenate with the normalized $\boldsymbol{M_i'}$ as input, following previous works [8, 14]. We employ a parallel, high-resolution structure for the feature extractor f_{ext}, inspired by the improvement achieved in the human pose estimation task [11]. f_{ext} is an n-multi-branch shared-weight feature extraction network, which can be expressed as follows:

$$\boldsymbol{\Phi}_i^{fr}, \boldsymbol{\Phi}_i^{hr}, \boldsymbol{\Phi}_i^{qr} = f_{ext}(\boldsymbol{M'}_i, \boldsymbol{L}_i),\ i \in \{1, 2, \cdots, n\}, \tag{2}$$

where $\boldsymbol{\Phi}_i^{fr} \in \mathbb{R}^{H\times W\times 64}$, $\boldsymbol{\Phi}_i^{hr} \in \mathbb{R}^{\frac{1}{2}H\times\frac{1}{2}W\times 128}$, and $\boldsymbol{\Phi}_i^{qr}$ $\in \mathbb{R}^{\frac{1}{4}H\times\frac{1}{4}W\times 256}$ are features of three different scales (full-resolution, half-resolution, and quarter-resolution) extracted by f_{ext}. We employ HR-Net [11] for two reasons. First, the structure increases the resolution in a parallel manner, avoiding passing the input from high-resolution layers to low-resolution layers. Second, the maximum responses represent different features at different scales. Therefore, the multiscale features can provide more useful information for the subsequent modules.

Multiscale bilateral extraction & top-k pooling: As described in Sections 2.1 and 2.2, these two modules can fuse an arbitrary number of input features into a single feature with a fixed number of channel. Therefore, we employ bilateral extraction to handle the n input features at three scales, and then use top-k pooling on six groups (three scales × bilateral extraction) of features, for the next normal regression stage. In our method, we set the hyperparameter k = 3 and use ELU as the bilateral activation function, to achieve the best performance (see Section 3.1).

Normal regression: The aim of normal regression f_{reg} is to regress surface normals by the aggregated features ($\boldsymbol{\Psi}^{fr} \in \mathbb{R}^{H\times W\times 64k}$, $\boldsymbol{\Psi}^{hr} \in \mathbb{R}^{\frac{1}{2}H\times\frac{1}{2}W\times 128k}$, $\boldsymbol{\Psi}^{qr}$ $\in \mathbb{R}^{\frac{1}{4}H\times\frac{1}{4}W\times 256k}$), as follows:

$$\tilde{\boldsymbol{N}} = f_{reg}(\boldsymbol{\Psi}^{fr}, \boldsymbol{\Psi}^{hr}, \boldsymbol{\Psi}^{qr}), \tag{3}$$

where $\tilde{\boldsymbol{N}}$ is the estimated surface normal. We first employ the transposed convolution operations to up-sample the low-resolution features $\boldsymbol{\Psi}^{hr}$ and $\boldsymbol{\Psi}^{qr}$ to full resolution $H \times W$. Then, we concatenate the two up-sampled features with $\boldsymbol{\Psi}^{fr}$, followed by two convolutional layers to reduce the number of channels to three. Note that we employ L2 normalization at the end, to make the per-pixel estimated surface normal $\tilde{\boldsymbol{n}} \in \tilde{\boldsymbol{N}}$ become a unit vector.

We train HPS-Net by minimizing the angular error between the estimated and the ground-truth surface normals (*i.e.*, cosine similarity loss), as follows:

$$loss = \frac{1}{HW}\sum_{p}(1 - \boldsymbol{N_p} \odot \tilde{\boldsymbol{N}}_{\boldsymbol{p}}), \tag{4}$$

where $\odot$ represents the dot-product operation, $\boldsymbol{N_p}$ and $\tilde{\boldsymbol{N}}_{\boldsymbol{p}}$ are the ground-truth and predicted surface normals at the pixel with index p. HPS-Net is trained with two synthetic data sets, called the Blobby shape dataset [15] and the Sculpture shape dataset [16], with a total of 852 samples selected as the validation set, which is the same as most previous work, for a fair comparison. The proposed HPS-Net was implemented using PyTorch with Adam optimizer (β_1 = 0.9 and β_2 = 0.999). The initial learning rate is set to 0.002, divided by 2 every 5 epochs. We trained the model, using a batchsize of 16, for 40 epochs. The number of input images used for training is 32, and variable in testing. Furthermore, we set the resolution $H \times W$ to 32×32 for training.

3. EXPERIMENTS

In this section, we present the experiments and analysis for our proposed bilateral extraction, top-k pooling and HPS-Net. We evaluate the accuracy of the estimated surface normal using the mean angular error (MAE) in degrees, where MAE = $\frac{1}{HW}\sum_{p}^{H\times W}\cos^{-1}\left(\boldsymbol{N}_p \odot \tilde{\boldsymbol{N}}_p\right)$. A lower MAE indicates better estimation accuracy.

3.1. Ablation study

We select PS-FCN [8] as the baseline network, to test the effectiveness of the proposed bilateral extraction and top-k pooling.

First, we test the effect of the different kinds of activation functions (ReLU [17] and ELU [18]) on the baseline network, as tabulated in Table 1. Note that α is set to 1 in ELU. The activatio functions are evaluated in terms of MAE on the validation set, with 64 input images.

As tabulated in Table 1, the proposed bilateral extraction module can improve the estimation results, when compared to the normal activation functions. Furthermore, it can be seen that the ELU outperforms the ReLU activation function. This can be explained by the fact that the Elu [18] weakens the negative parts of features, providing partial information, while ReLU [17] directly truncates the negative part of all features. Therefore, the rotated negative feature generated by ELU retains more information. However, these two activation functions are the same for the positive parts of features (*i.e.*, without using bilateral extraction).

Table 1. The performance of different activation functions, with and without bilateral extraction, in terms of MAE.

Activation functions	MAE
ReLU ($w/$ bilateral)	6.46
ReLU (w/o bilateral)	6.59
ELU ($w/$ bilateral)	**6.39**
ELU (w/o bilateral)	6.57

Table 2. The performance of using different feature fusion methods, in terms of MAE.

Manners	Testing number			
	10	16	32	64
1 × 1 Conv.	8.74	8.05	-	-
Max-pooling	8.62	7.92	7.27	6.59
Top-k pooling (k =2)	8.57	7.84	7.13	6.52
Top-k pooling (k =3)	**8.54**	**7.79**	7.04	6.44
Top-k pooling (k =4)	8.58	7.80	7.02	6.41
Top-k pooling (k =5)	8.65	7.82	**7.01**	**6.40**

Furthermore, we test the effectiveness of top-k pooling. In Table 2, we compare the results of top-k pooling (using different k), max-pooling, and 1 × 1 convolutional layer, on the validation set. It is worth noting that a 1 × 1 convolutional layer needs a fixed input dimension. Therefore, we trained the baseline network (PS-FCN) with 10, 16 inputs (the maximum number of input images of the validation set and training set is 64, and the GPU memory fails to fuse a feature with 128 × 32 and 128 × 64 channels). For top-k pooling and max-pooling, the network is trained with the same 32 input images.

As shown in Table 2, we can see that top-k pooling consistently outperforms max-pooling and 1 × 1 convolutional layer. This demonstrates the effectiveness of the top-k pooling method. In fact, the selection of the hyperparameter k is important for our method. A larger k (*e.g.*, 4, 5) may lead to worse results than a smaller k under sparse input conditions. This is because a larger k in sparse input trends to represent the average information that will smooth salience, while still being the high-responded information under dense input condition. However, a larger k (*e.g.*, more than 3) only brings a slight improvement, but increases the computational burden. Therefore, we choose k = 3 in our HPS-Net.

3.2. Benchmark comparisons

The widely used DiLiGenT benchmark [19] contains 10 objects of different shapes and complex non-Lambertian surfaces. In Table 3, we first show the performance of other max-pooling methods replaced by our proposed bilateral extraction and top-k modules (B&T), with 96 input images. Then, the test results of our HPS-Net and other state-of-the-art methods on the DiLiGenT benchmark [19] are listed in Table 4. We also illustrate the visual results in Fig. 3.

As tabulated in Table 3, the proposed bilateral extraction and top-k modules can improve the performance of the three max-pooling-based methods. This is attributed to the fact that bilateral extraction additionally brings the truncated negative features, and top-k pooling preserves k times more information than traditional max-pooling fusion. Moreover, as shown in Table 4, our HPS-Net achieves the advanced MAE for complex objects, such as “Pot2”, “Reading”, and those strong non-Lambertian objects, such as “Goblet”, “Cow”, which contain shadows and inter-reflections, as shown in Fig. 3. Overall, the proposed method achieves superior performance.

Table 3. The performance of the average MAE on the DiLiGenT benchmark [19]. B&T represents the proposed bilateral extraction and top-k.

w/o B&T	PS-FCN [8]	Attention-PSN [14]	PS-FCN(N.) [12]
	8.39	7.92	7.39
$w/$ B&T	PS-FCN [8]	Attention-PSN [14]	PS-FCN(N.) [12]
	8.08	**7.72**	**7.14**

Table 4. Comparison of different methods on the DiLiGenT benchmark [19]. All methods are evaluated with 96 images.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
LS [1]	4.10	8.39	14.92	8.41	25.60	18.50	30.62	8.89	14.65	19.80	15.39
DPSN [20]	2.02	6.31	12.68	6.54	8.01	11.28	16.86	7.05	7.86	15.51	9.41
IRPS [21]	1.47	5.79	10.36	5.44	6.32	11.47	22.59	6.09	7.76	11.03	8.83
LMPS [22]	2.40	5.23	9.89	6.11	7.98	8.61	16.18	6.54	7.48	13.68	8.41
PS-FCN [8]	2.82	7.55	7.91	6.16	7.33	8.60	15.85	7.13	7.25	13.33	8.39
Attention-PSN[14]	2.93	4.86	7.75	6.14	6.86	8.42	15.44	6.92	6.97	12.90	7.92
DR-PSN [23]	2.27	5.46	7.84	5.42	7.01	8.49	15.40	7.08	7.21	12.74	7.90
CHR-PSN [24]	2.26	6.35	7.15	5.97	6.05	8.32	15.32	7.04	6.76	12.52	7.77
CNN-PS [7]	2.12	8.30	8.07	4.38	7.92	7.42	14.08	5.37	6.38	12.12	7.62
GPS-Net [25]	2.92	5.07	7.77	5.42	6.14	9.00	15.14	6.04	7.01	13.58	7.81
PS-FCN(N.) [12]	2.67	7.72	7.53	4.76	6.72	7.84	12.39	6.17	7.15	10.92	7.39
MF-PSN [26]	2.07	5.83	6.88	5.00	5.90	7.46	13.38	7.20	6.81	12.20	7.27
HPS-Net (our)	2.37	5.28	6.89	4.98	5.59	7.59	14.17	6.23	6.77	11.26	**7.11**

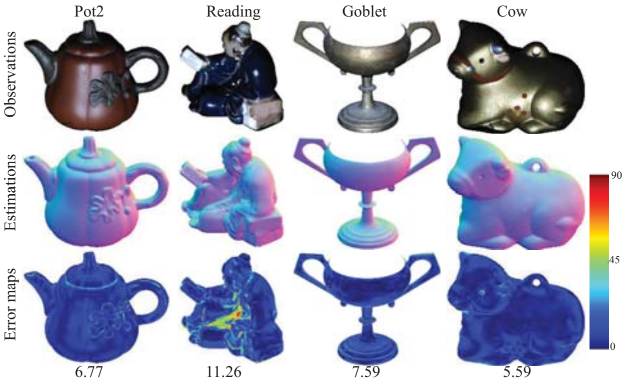

Fig. 3. Visual results of the proposed HPS-Net, on the DiLiGenT benchmark [19], with 96 input images.

4. CONCLUSIONS

In this paper, we proposed a deep model for efficient and accurate photometric stereo, namely HPS-Net, based on our proposed bilateral extraction module and top-k pooling module. These two modules can better solve the fusion problem of handling an arbitrary number of input features in learning-based photometric stereo. Ablation studies demonstrate that the bilateral extraction module can achieve more accurate surface-normal prediction than using the conventional activation functions and top-k pooling outperforms max-pooling and 1 × 1 convolution. Experiment results on the DiLiGenT benchmark show the effectiveness of the proposed modules. Furthermore, the proposed HPS-Net shows superior performance compared to other deep learning-based photometric stereo networks.

5. REFERENCES

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

[2] Lixiong Chen, Yinqiang Zheng, Boxin Shi, Art Subpa-Asa, and Imari Sato, "A microfacet-based reflectance model for photometric stereo with highly specular surfaces," in *ICCV*, 2017, pp. 3162–3170.

[3] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa, "Robust photometric stereo using sparse regression," in *CVPR*, 2012, pp. 318–325.

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

[5] Jun Xiao, Qian Ye, Rui Zhao, Kin-Man Lam, and Kao Wan, "Self-feature learning: An efficient deep lightweight network for image super-resolution," in *ACM MM*, 2021, pp. 4408–4416.

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

[7] Satoshi Ikehata, "Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces," in *ECCV*, 2018, pp. 3–18.

[8] Guanying Chen, Kai Han, and Kwan-Yee K Wong, "Ps-fcn: A flexible learning framework for photometric stereo," in *ECCV*, 2018, pp. 3–18.

[9] Yakun Ju, Kin-Man Lam, Wuyuan Xie, Huiyu Zhou, Junyu Dong, and Boxin Shi, "Deep learning methods for calibrated photometric stereo and beyond: A survey," *arXiv preprint arXiv:2212.08414*, 2022.

[10] Zi-Rong Jin, Liang-Jian Deng, Tian-Jing Zhang, and Xiao-Xu Jin, "Bam: Bilateral activation mechanism for image fusion," in *ACMMM*, 2021, pp. 4315–4323.

[11] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

[12] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee Kenneth Wong, "Deep photometric stereo for non-lambertian surfaces," *TPAMI*, 2020.

[13] Imari Sato, Takahiro Okabe, Qiong Yu, and Yoichi Sato, "Shape reconstruction based on similarity in radiance changes under varying illumination," in *ICCV*, 2007, pp. 1–8.

[14] Yakun Ju, Kin-Man Lam, Yang Chen, Lin Qi, and Junyu Dong, "Pay attention to devils: A photometric stereo network for better details," in *IJCAI*, 2020, pp. 694–700.

[15] Micah K Johnson and Edward H Adelson, "Shape estimation in natural illumination," in *CVPR*, 2011, pp. 2553–2560.

[16] Olivia Wiles and Andrew Zisserman, "Silnet: Single- and multi-view reconstruction by learning from silhouettes," in *BMVC*, 2017.

[17] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[18] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *ICLR*, 2016.

[19] 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.," *TPAMI*, vol. 41, no. 2, pp. 271–284, 2019.

[20] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita, "Deep photometric stereo network," in *ICCVW*, 2017, pp. 501–509.

[21] Tatsunori Taniai and Takanori Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *ICML*, 2018, pp. 4857–4866.

[22] Junxuan Li, Antonio Robles-Kelly, Shaodi You, and Yasuyuki Matsushita, "Learning to minify photometric stereo," in *CVPR*, 2019, pp. 7568–7576.

[23] Yakun Ju, Junyu Dong, and Sheng Chen, "Recovering surface normal and arbitrary images: A dual regression network for photometric stereo," *TIP*, vol. 30, pp. 3676–3690, 2021.

[24] Yakun Ju, Yuxin Peng, Muwei Jian, Feng Gao, and Junyu Dong, "Learning conditional photometric stereo with high-resolution features," *Computational Visual Media*, vol. 8, no. 1, pp. 105–118, 2022.

[25] Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, and Boxin Shi, "Gps-net: Graph-based photometric stereo network," in *NeurIPS*, 2020.

[26] Yanru Liu, Yakun Ju, Muwei Jian, Feng Gao, Yuan Rao, Yeqi Hu, and Junyu Dong, "A deep-shallow and global–local multi-feature fusion network for photometric stereo," *Image and Vision Computing*, vol. 118, pp. 104368, 2022.