PS-Transformer: Learning Sparse Photometric Stereo Network using Self-Attention Mechanism

Satoshi Ikehata
https://satoshi-ikehata.github.io/

National Institute of Informatics
Tokyo, JAPAN

Abstract

Existing deep calibrated photometric stereo networks basically aggregate observations under different lights based on the pre-defined operations such as linear projection and max pooling. While they are effective with the dense capture, simple first-order operations often fail to capture the high-order interactions among observations under small number of different lights. To tackle this issue, this paper presents a deep sparse calibrated photometric stereo network named *PS-Transformer* which leverages the learnable self-attention mechanism to properly capture the complex inter-image interactions. PS-Transformer builds upon the dual-branch design to explore both pixel-wise and image-wise features and individual feature is trained with the intermediate surface normal supervision to maximize geometric feasibility. A new synthetic dataset named CyclesPS+ is also presented with the comprehensive analysis to successfully train the photometric stereo networks. Extensive results on the publicly available benchmark datasets demonstrate that the surface normal prediction accuracy of the proposed method significantly outperforms other state-of-the-art algorithms with the same number of input images and is even comparable to that of dense algorithms which input $10\times$ larger number of images.

1 Introduction

Photometric Stereo is a long-standing problem to recover a fine surface normal map from HDR (High Dynamic Range) images captured under different lights with a fixed camera. Since Woodham [27] proposed the first Lambertian calibrated photometric stereo algorithm, optimization based inverse rendering had been a mainstream approach [1, 9, 15, 16, 23]. However, recent breakthroughs from deep neural networks have sparked great interest to explore the data-driven photometric stereo algorithms to handle complex global illumination effects that cannot be described in a mathematically tractable form.

Unlike most computer vision tasks, a photometric stereo network must accept a *set input* [17] (*i.e.*, unordered, varying number of pairs of image and light) and the output should take into account the entire input set. Existing deep photometric stereo networks satisfied this requirement by mainly two permutation invariant aggregation methods, which are *observation-map* [13, 18, 32] and *set-pooling* [5, 6, 7]. The former introduces the fixed-shape 2-d map called *observation map* where all the observations at a single pixel are projected

according to light directions to be fed to convolutional neural networks (CNN) for pixel-wise surface normal prediction. On the other hand, the latter is based on the set-pooling method [30] where a feature map is firstly encoded from a pair of image and light, then all the feature maps from different images are merged using a pooling operation (*e.g.*, mean, max) to be fed to CNN to predict the 2-d surface normal map.

While both approaches accept a set input of varying size, recent follow-up works [18, 29, 32] which compared these two approaches pointed their drawbacks. Given enough input images, an observation map captures shading variations of individual pixels better than set-pooling algorithms, however its performance significantly drops in the *sparse* photometric stereo setup (*i.e.*, the number of input images is small e.g., less than 10) where shading variations of individual pixels are not sufficient to recover surface details. Since photometric stereo data acquisition requires large labor, it is more convenient to recover accurate normal map with the least number of images and the set-pooling algorithms have an advantage here owing to the feature map which encodes the intra-image shading information. To get the best of both approaches, Yao *et al*. [29] have recently proposed a two-step approach named *GPS-Net*, which firstly aggregated the per-pixel shading variations using the trainable structure-aware graph convolution (SGC) [4], then constructed a feature map by merging features of different pixels to be fed to CNN-based surface normal predictor to account for the intra-image spatial information. However, there are two problems in this method. First, their SGC filters constructed the neighborhood structure only from the virtual central node to each observation of the same pixel therefore incorporated only the first-order proximity. More recently, Liu *et al*. [20] have proposed *SPS-Net* which firstly introduced the self-attention mechanism [26] to encode higher-order interactions among observations under different lights. The self-attention does not assume the fixed-size input nor regular grid structure therefore doesn't encounter the sparsity problem in the observation map. SPS-Net repeats a set of the self-attention block and some convolutions repeatedly with different image scales and finally performs the max-pooling to feed the features to the CNN-based normal map predictor. While the self-attention mechanism in SPS-Net may capture high-order interactions among observations, the intra-image spatial feature and inter-image photometric feature were intermediately jumbled up in the model via multiple convolutions and accompanied downsampling, therefore important perpixel shading information was not maximally utilized. In addition, SPS-Net aggregated all the information via the max-pooling at last, which could squash the higher order information before the final normal map prediction.

Based on these insights, this paper presents a transformer-based photometric stereo network, namely *PS-Transformer* that overcomes limitations in existing deep photometric stereo networks. The main ideas are two-fold. First, as with [20], the self-attention mechanism in the transformer model [26] is introduced to aggregate features under different lights, however, multiple, multi-head self-attention layers are stacked in a row rather than alternating the self-attention and convolutions as in [20] to keep position specific information. Second, we introduce the dual-branch design to discriminate inter-image and intra-image information explicitly at the feature aggregation phase and fuse them just before the final prediction. In addition, the new synthetic training dataset, namely *CyclesPS+*, is also presented with training strategy optimized for it. It will be shown that PS-Transformer trained on CyclesPS+ dataset could dramatically improve the performance especially in the sparse photometric stereo problem, which is almost comparable to the performance on the dense input set.

Preliminaries: The goal of the calibrated photometric stereo is to recover a surface normal map ($N \in \mathbb{R}^{h\times w\times 3}$) from images ($I_1, \cdots, I_m \in \mathbb{R}^{h\times w\times c}$) captured under known directed

lights ($\boldsymbol{l}_1, \cdots, \boldsymbol{l}_m | \boldsymbol{l}_j = [l_j^x, l_j^y, l_j^z] \in \mathbb{R}^3$) with a known fixed camera ($\boldsymbol{v} = [0,0,1]^\top$) where w, h and c are width, height and color channel of the image and m is the total number of different lights. Henceforth j indicates the light index and i indicates the pixel index (*e.g.*, $\boldsymbol{x}_{j,i}$ is the value at i-th pixel of j-th matrix $\boldsymbol{x}_j$). We assume that pixel intensities are normalized by their power of light, therefore $\boldsymbol{l}$ is supposed to be a unit vector.

2 Related Works

2.1 Conventional Photometric Stereo Algorithms

Before the advent of deep learning, most conventional photometric stereo algorithms recovered surface normals of a scene via a simple diffuse reflectance modeling (*e.g.*, Lambertian) while treating other effects as outliers [15, 16, 22, 28]. While effective, a drawback of this approach is that if it were not for dense diffuse inliers, the estimation fails. To handle dense non-Lambertian reflections, various algorithms arrange the parametric or non-parametric models of non-Lambertian BRDF [2, 9, 14, 23]. Even though nonlinear models can be applied to a variety of materials, they were powerless against model outliers. A few amount of photometric stereo algorithms are grouped into the example-based approach, which takes advantages of the surface reflectance of objects with known shape, captured under the same illumination environment with the target scene [12, 25]. While effective, this approach also suffers from model outliers and has a drawback that the lighting configuration of the reference scene must be taken over at the target scene.

2.2 Deep Photometric Stereo Algorithms

Deep photometric stereo networks basically consist of two modules (a) permutation invariant feature aggregation module and (b) surface normal prediction module. Given a set of input images and corresponding lights, the permutation invariant feature aggregation module encodes pixel-wise features as

$$\boldsymbol{f}_i = \mathrm{Agg}\{\boldsymbol{x}_{1,i}, \cdots, \boldsymbol{x}_{m,i}\}. \tag{1}$$

Here, $\boldsymbol{f}_i$ is a d_{agg}-dimensional vector (d_{agg} varies by algorithm) and $\boldsymbol{x}_{j,i}$ is the feature from i-th pixel under j-th light and "Agg" is the aggregation over features at the same pixel under different lights in the permutation invariant manner. Both the aggregation operation and feature type differ between algorithms. In the observation-map based algorithm [13, 18, 32], $\boldsymbol{x}_{j,i} \triangleq [I_{j,i}, \boldsymbol{l}_j]$ and the aggregation operation is a set of projections of $I_{j,i}$ onto l_x-l_y coordinates of the fixed size observation map. On the other hand, the set-pooling based algorithms [5, 6, 7] define $\boldsymbol{x}_{j,i} \triangleq \phi(I_j, \boldsymbol{l}_j)_i$ where $\phi(I_j, \boldsymbol{l}_j)$ is the feature map extracted from the image and light $\{I_j, \boldsymbol{l}_j\}$ and the aggregation is pixel-wise max or mean pooling of feature maps. GPS-Net [29] also defines $\boldsymbol{x}_{j,i} \triangleq [I_{j,i}, \boldsymbol{l}_j]$ but the aggregation is based on the trainable SGC filters. Given aggregated features, the surface normal is predicted by either form of $N_i = \psi(\boldsymbol{f}_i)$ or $N = \Psi(\boldsymbol{f}_1, \cdots, \boldsymbol{f}_{wh})$. Here, ψ is the pixel-wise surface normal predictor (*i.e.*, a feed-forward neural network) and Ψ is the image-wise surface normal predictor (*i.e.*, a convolutional neural network) where individual features merged into a single feature map before the prediction.

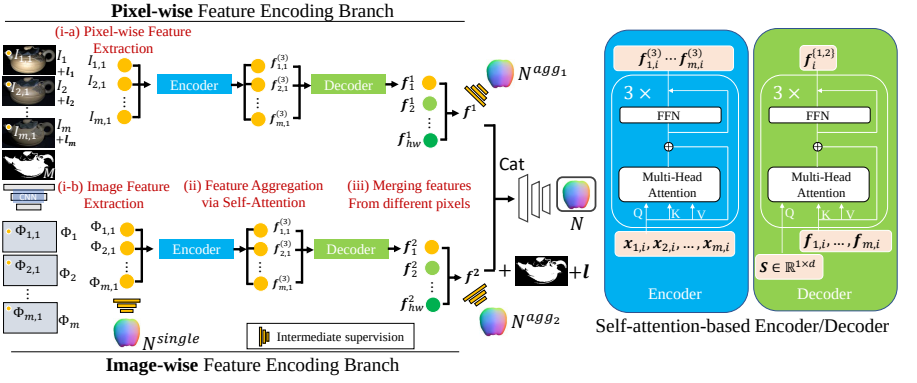

Figure 1: The illustration of the PS-Transformer. Given m number of input images with corresponding lights, the input is (i-a) directly passed to the pixel-wise feature encoding module for taking advantage of pixel-wise shading variations and also passed to the image-wise feature encoding module (i-b) after the feature maps are extracted from each image independently for taking advantage of spatial information. In each module, (ii) per-pixel feature vectors under different lights are aggregated by the encoder and the decoder based on the multi-head self-attention blocks and (iii) the aggregated feature vectors are merged into the single feature map by another self-attention layer to be fed to the CNN-based surface normal map predictor. To link intermediate features with surface normal information, the intermediate surface normal supervision is introduced when training the network.

3 PS-Transformer

In this section, we introduce our transformer-based photometric stereo network, called *PS-Transformer*. The entire architecture is illustrated in Fig. 1.

3.1 Feature Aggregation using Self-Attention

We first describe the feature aggregation operation (*i.e.*, Agg in Eq. (1)) in PS-Transformer which is the core component in our framework. Similar to other transformer-based architectures, our feature aggregation consists of an encoder network (*i.e.*, concurrently encode the whole set) followed by a decoder network (*i.e.*, pooling encoded features) as follow

$$\boldsymbol{f}_i = \text{Decoder}(\text{Encoder}\{\boldsymbol{x}_{1,\text{i}}, \cdots, \boldsymbol{x}_{\text{m,i}}\}). \tag{2}$$

Before digging into the details, we first review the self-attention layer in the transformer model [26]. Given a set of features $(\boldsymbol{x}_j | 1 \leq j \leq m)$, the self-attention layer firstly projects each item onto three different vectors: the query vector $W^Q \boldsymbol{x}_j$, the key vector $W^K \boldsymbol{x}_j$ and the value vector $W^V \boldsymbol{x}_j$ with embedding dimension d_q, d_k, d_v. Vectors computed from different items are then packed together into three different matrices, namely, $Q \in \mathbb{R}^{m \times d_q}$, $K \in \mathbb{R}^{m \times d_k}$ and $V \in \mathbb{R}^{m \times d_v}$. The output $A(Q,K,V) \in \mathbb{R}^{m \times d_v}$ of the self-attention layer is given by,

$$A(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} V\right). \tag{3}$$

The self-attention layer computes the attention scores as normalized dot-product of the query with all keys and outputs the weighted sum of values where a value gets more weight if its corresponding key is similar to the query. Note that as derived from Eq. (3), the self-attention layer is invariant to permutations and changes in the number of input items. Since a single-head self-attention layer limits its ability to focus on simple interaction among items, it is common to comprises multiple self-attention blocks (*i.e.*, multi-head attention). In this work, we also use a h-head multi-head attention layer ($h = 8$ in our evaluation) and stack multiple multi-head attention layers in a row to encode higher order interactions.

Using this self-attention layer, Encoder(.) in Eq. (2) is represented by three sets of multi-head attention and feed-forward networks (FFN) as

$$\begin{aligned} F_i^{(t)} &\triangleq [\boldsymbol{f}_{0,i}^{(t)} \cdots \boldsymbol{f}_{m,i}^{(t)}]^\top \in \mathbb{R}^{m\times d} \\ Q &= F_i^{(t)} W^Q \in \mathbb{R}^{m\times d_q},\ K = F_i^{(t)} W^K \in \mathbb{R}^{m\times d_k},\ V = F_i^{(t)} W^V \in \mathbb{R}^{m\times d_v}, \\ H &= Q + \mathrm{MultiheadAttn}(Q,K,V), \\ F_i^{(t+1)} &= \mathrm{FFN}(\mathrm{GeLU}(\mathrm{FFN}(H))) + H, \end{aligned} \tag{4}$$

where $\boldsymbol{f}_{j,i}^{(t)} \in \mathbb{R}^d$ is the feature vector at i-th pixel under j-th light (*i.e.*, $\boldsymbol{f}_{j,i}^{(0)} = \boldsymbol{x}_{j,i}$) and W^Q, W^K and W^V are the projection matrices and GeLU is the Gaussian error Linear Units [11]. Note that we use the dropout (the probablity is 0.1) after the activation but does not use the Layer Normalization (LayerNorm) which is commonly used in the exisiting transformer models because we empirically found it degrades the performance. The dimensionality of the hidden layers is fixed by $d = d_q = d_k = d_v = 256$ throughout the manuscript.

Given the input features $\{\boldsymbol{x}_{1,i}, \cdots, \boldsymbol{x}_{m,i}\}$, the output of the encoder is a set of encoded features contained in the matrix $F_i^{(3)} = [\boldsymbol{f}_{1,i}^{(3)} \cdots \boldsymbol{f}_{m,i}^{(3)}]^\top \in \mathbb{R}^{m\times d}$. For aggregating the information over different lights, we then apply the Decoder(.) in Eq. (2) to shrink the feature size without losing the interactions among set items as is independent of the input set size m. To achieve this, we introduce the PMA (Pooling by Multihead Attention) module [17] which applies multi-head attention on a learnable seed vector $S \in \mathbb{R}^{1\times d}$ as key vector and use $F_i^{(3)}$ as query and value vectors to get perpixel d-dimensional vectors $\boldsymbol{f}_i$. The next section details our photometric stereo network architecture that uses this feature aggregation operation.

3.2 Network Architecture

As have already been discussed in previous works [18, 29, 32], the intensity variations of individual pixels over different lights contribute to recover the sharp geometric boundaries and local spatial intensity variations among neighbor pixels contribute to recover geometry from a sparse input set. To get the best of them, PS-Transformer introduces the dual-branch design as depicted in Fig. 1. Both branches consist of self-attention based feature aggregation as detailed and the only difference in two branches is the type of input features to be aggregated. Concretely, the first branch takes $\boldsymbol{x}_{j,i}^1 \triangleq [I_{j,i}, \boldsymbol{l}_j] \in \mathbb{R}^{c+3}$ and the second branch takes $\boldsymbol{x}_{j,i}^2 \triangleq [\phi(I_j, M)_i, \boldsymbol{l}_j] \in \mathbb{R}^{67}$ as input where $M \in \mathbb{R}^{h\times w}$ is the object mask which gives the network the boundary constraint as is known to be helpful in the shape-from-shading literature [31][1]. ϕ is shared-weight convolutional neural networks with six 3×3 convolution layers, normalization (Batch Normalization) and activation (Leaky-ReLU) to output the sixty

[1]Object mask is simply acquired by taking the non-zero (or more than some small value for the robustness) pixels in the image averaged over all the light directions.

four dimensional feature map. The acquired feature map is then concatenated with a lighting map where $\boldsymbol{l}_j \in \mathbb{R}^3$ is expanded to the size of the image. Note that the encoder/decoder in two branches don't share the network parameters. The output from these two branches are concatenated with the object mask, then a feature map of $\mathbb{R}^{h\times w\times(2d+1)}$ is fed to image-space surface normal predictor Ψ which consists of five 3×3 convolution layers whose number of filters are $2d+1$ except for the final surface normal prediction layer to get $N \in \mathbb{R}^{h\times w\times 3}$. In summary, our PS-Transformer architecture is formally described as

$$\begin{aligned} \boldsymbol{f}_i^1 &= \text{Decoder}_1(\text{Encoder}_1\{\boldsymbol{x}_{1,i}^1,\cdots,\boldsymbol{x}_{m,i}^1\}), \quad \boldsymbol{x}_{j,i}^1 \triangleq [I_{j,i}, \boldsymbol{l}_j], \\ \boldsymbol{f}_i^2 &= \text{Decoder}_2(\text{Encoder}_2\{\boldsymbol{x}_{1,i}^2,\cdots,\boldsymbol{x}_{m,i}^2\}), \quad \boldsymbol{x}_{j,i}^2 \triangleq [\phi(I_j, M)_i, \boldsymbol{l}_j], \\ N &= \Psi(\text{cat}\{\boldsymbol{f}_1^1, \boldsymbol{f}_1^2, M_1\},\cdots,\text{cat}\{\boldsymbol{f}_{hw}^1, \boldsymbol{f}_{hw}^2, M_{hw}\}). \end{aligned} \tag{5}$$

where "cat" is the operation for the pixel-wise concatenation.

3.3 Loss Function

The network is trained with a simple mean squared loss between predicted and ground truth surface normal maps. In addition to evaluating the final output (N), we also put intermediate surface normal supervision to encourage each feature to directly associate with the surface normal prediction. Concrete form of the loss function is as follow,

$$L_{PST} = \|M\odot(N-N^{gt})\|_2 + \frac{1}{m}\sum_{j=1}^{m}\|M\odot(N_j^{single}-N^{gt})\|_2 \tag{6}$$

$$+ \|M\odot(N^{agg_1}-N^{gt})\|_2 + \|M\odot(N^{agg_2}-N^{gt})\|_2, \tag{7}$$

where $\odot$ denotes Hadamard product to remove background pixels from the loss computation. Here, N_j^{single} is the surface normal map predicted from the j-th single-view feature map $\boldsymbol{x}_j^2 \in \mathbb{R}^{h\times w\times 67}$ in Eq. (5) by six 3×3 convolution layers whose number of filters is 67 except for the final output layer with normalization (Batch Normalization) and activation (Leaky-ReLU). $N^{agg_{\{1,2\}}}$ are surface normal maps formed by predicted surface normal vectors from aggregated features ($\boldsymbol{f}_i^1$ and $\boldsymbol{f}_i^2$ $|1\le i\le hw$) by two fully-connected layers (the dimension changes as $d\to d\to 3$) with activation (Learky-ReLU) to predict three dimensional normal vector. We assigned equal weight to the contribution of each term without valuing specific normal map strongly.

4 Results

Algorithms: PS-Transformer is evaluated with representative photometric stereo networks based on the observation map (CNN-PS [13]), set-pooling (PS-FCN+ [7])[2] and the graph convolution (GPS-Net [29]). We also compared our method against *SPS-Net* [20] where the self-attention mechanism is also applied to interact features under different lights.

In our experiments, we used authors' official implementations and pretrained models [8, 10, 21] with minor modifications to evaluate them with the same training/test protocol [3].

[2] PS-FCN+ [7] is the extension from PS-FCN [5] where data normalization strategy to equalize spatial appearance has been introduced.

[3] Please see supplementary materials for further details.

For the fair comparison, we also compared against existing models trained on our training data which results in five competitors in total; (i) PS-FCN+ (Trained on our dataset), (ii) PS-FCN+ (Pretrained), (iii) GPS-Net (Trained on our dataset), (iv) GPS-Net (Pretrained), (v) CNN-PS (Trained on our dataset)[4] and (iv) SPS-Net (Trained on our dataset). The original implementation of SPS-Net was trained on 32×32 patches of the Blobby and Sculpture datasets [5], however it is not clear if this patch size is optimal for our CyclesPS+ dataset which will be described later. Therefore we compared two different configurations of SPS-Net where the network was trained on either of 32×32 or 8×8 (same training patch size as PS-Transformer) patches for the fair comparison.

Training Dataset: The existing deep photometric stereo models were trained either on Blobby shape datasets [5] or CyclesPS datasets [13]. The Blobby shape dataset contains much bigger number of samples (*i.e.*, 85212), however the material is spatially uniform and lighting variations in each dataset is small (*i.e.*, 64). On the other hand, CyclesPS datasets only contain 15 different objects but the material is spatially varying and the lighting variations in each dataset is large (*i.e.*, 740). To get the best of them, we created *CyclesPS+* dataset following the rendering scheme in CyclesPS but increased the dataset size (25 objects including 15 objects in CyclesPS and different types of subsets for each object). Please refer supplementary materials for further instruction.

Training Details: Since we mainly target the sparse photometric stereo problem (*i.e.*, $m \leq 10$), the number of lights in a training sample is always fixed by 10 for training all the models including ours and competitors (except when using pretrained models). We should note that it has often been reported that the test accuracy improves when the number of images at the training phase matches with one for test [7]. However, it is quite inefficient to have models for every different number of images, therefore we reuse a single trained model (*i.e.*, $m = 10$) for a varying size of input set at test. Due to the space limit, the detailed training protocol (*e.g.*, number of epochs or learning strategy) is presented in the supplementary.

Test Dataset: Our main result is based on DiLiGenT [24] and DiLiGenT-MV [19]. The number of real objects in total is 110 (10 objects $\times$ 1 view [24] and 5 objects $\times$ 20 views [19]) which is significantly larger than all the existing evaluations in [7, 13, 29] where only DiLiGenT [24] was used. Each data provides 16-bit integer HDR images with a resolution of 612×512 from 96 different known lighting directions. The ground truth surface normals for the orthographic projection and the single-view setup are also provided. The inference of models for sparse photometric stereo is basically unstable according to the light distribution (*i.e.*, the condition number is large), therefore we performed 10 random trials and averaged them. Specifically, ten sets of m random light directions were sampled from the upper hemispherical surface in advance, and *the same sets of the light distribution were used for all methods*. .[5] We also note that we didn't include the results on synthetic data in our work, however important analysis on synthetic data such as the study about the ability of Principled BRDF [3] in representing the real materials have already been provided in the previous work [13].

[4] We didn't use the pretrained model of CNN-PS since the public CNN-PS pretrained model was trained on large number of input images (*i.e.*, $m \geq 30$). As has already been mentioned in [7], when the number of images between training and test is largely different, the performance significantly drops.

[5] The effect of the different light distribution is discussed in the supplementary material.

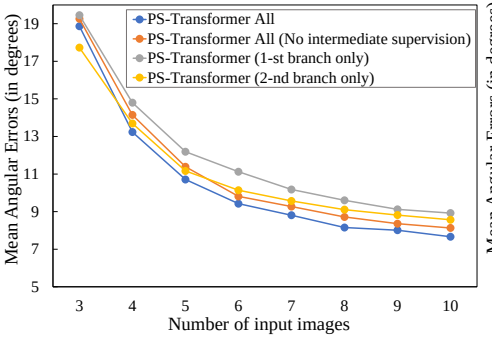

Figure 2: Ablation analysis to justify the design of our architecture. Errors for 10 objects are averaged.

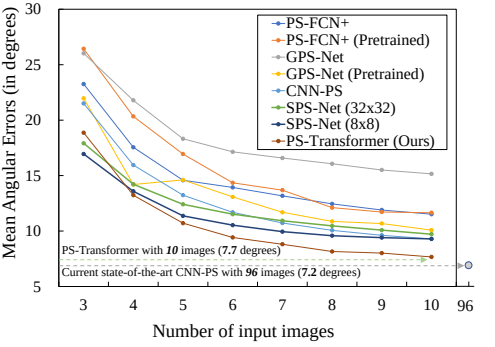

Figure 3: The quantitative comparison on DiLiGenT dataset. Errors for 10 objects are averaged.

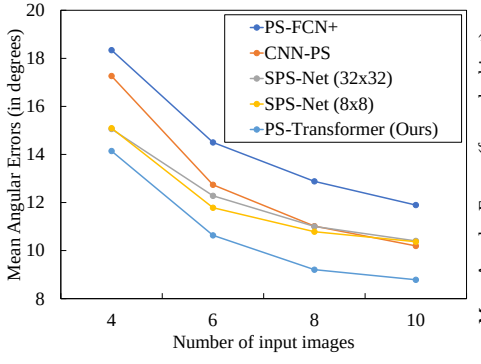

Figure 4: The quantitative comparison on DiLiGenT-MV dataset. Errors for five objects and twenty views are averaged.

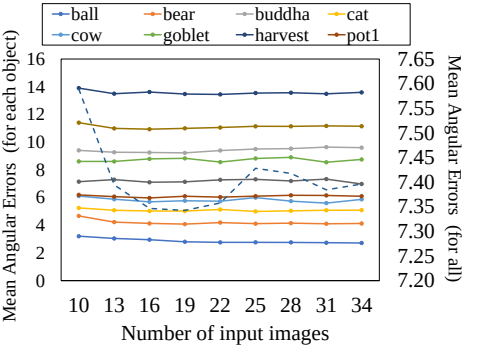

Figure 5: The sutdy with test images whose number is more than 10.

Ablation Study: We give ablation analysis to justify the design choice of PS-Transformer model including dual-branch design and the intermediate surface normal supervision. Here, we compared four different architectures on DiLiGenT dataset; (a) the architecture with a complete set of features (All) and (b) the same architecture but without the intermediate supervision, (c) the architecture only with the first branch (remove $\boldsymbol{f}^2$ from surface normal prediction module but with intermediate supervision), and (d) the architecture only with the second branch (remove $\boldsymbol{f}^1$ from surface normal prediction module but with intermediate supervision). The result is shown in Fig. 2. Here, we showed the mean angular errors averaged over 10 DiLiGenT objects. We observed that PS-Transformer model with the dual-branch architecture showed the best performance as expected. It was also observed that the intermediate supervision improved the prediction accuracy.

Quantitative Evaluation on DiLiGenT Dataset: We compared our method (full configuration) against five competitors mentioned in "Algorithms" on DiLiGenT main dataset in

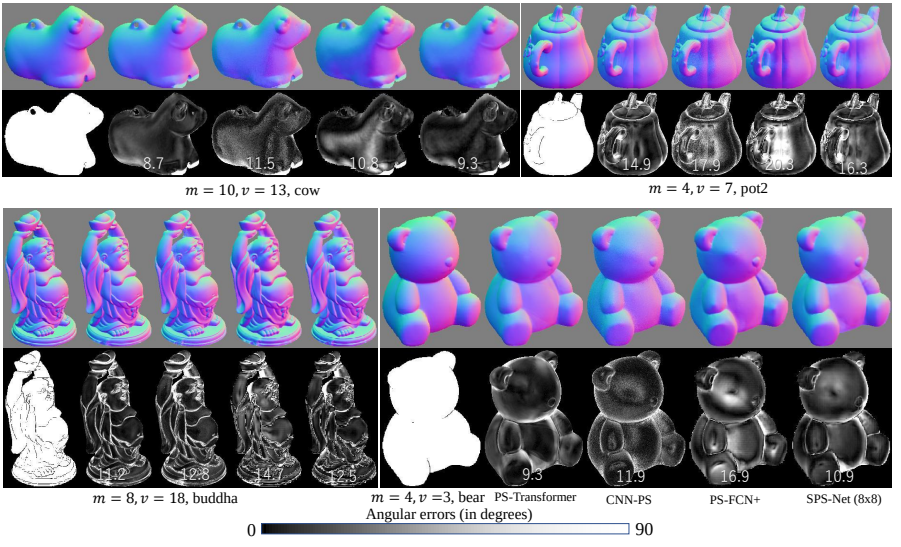

Figure 6: Qualitative evaluation on DiLiGenT-MV dataset.

the sparse photometric stereo setup ($3 \leq m \leq 10$). The results are illustrated in Fig. 3. Here, the mean angular errors (MAE) averaged over 10 objects are presented[6]. First, we observe that our method consistently outperforms other competitors in most cases. When the number of images is 10, the accuracy is comparable to the current state-of-the-art *dense* photometric stereo algorithm (*i.e.*, MAE = 7.2 with 96 images as reported in [13]) which shows the effectiveness of our transformer-based aggregation and dual-branch design. An interesting observation is that purely perpixel CNN-PS achieved roughly second best performance, which is even better than GPS-Net. This is contrary to the results in recent works [18, 29, 32] which claimed that a naïve observation map does not work well for the sparse photometric stereo setup.

SPS-Net [20] also introduces the self-attention mechanism for interacting information under different lights, however this method has inferior performance to our method. One of the obvious reasons could be that SPS-Net inherit the image-wise feature maps extracted at very early stage to the end therefore less took advantage of the per-pixel shading variations. In addition, SPS-Net doesn't introduce the intermediate supervision where our ablation study has already proved that it significantly contributes to improve the performance. There are also many differences in the design of the architecture. SPS-Net repeats the *single* self-attention block (called PF-Block in [20]) and the convolution repeatedly with different image scales while our method stacks the multiple, multi-head self-attention in a row *without* changing the image scale which contributes to keep the shape surface details. Furthermore, SPS-Net simply performs the max-pooling operation after the PF-Block for aggregating the feature maps under different lights unlike ours using PMA module [17] for decoding feature maps in more data-adaptive manner. It is interesting to observe that the performance of SPS-Net (8×8) consistently outperforms SPS-Net (32×32) because it indicates that the photometric stereo networks work better when observing the local shading variations rather

[6]Please refer supplementary material to see prediction errors of individual objects.

than observing more global information. This also supports the advantages of our dual-branch design that considers both local and global information.

Quantitative Evaluation on DiLiGenT-MV Dataset: The major drawbacks in DiLiGenT dataset is that the number of objects is only 10 and the variation of the surface normal distribution was quite limited. Recently, DiLiGenT dataset was extended to the multi-view edition as DiLiGenT-MV [19]. This new dataset contains images of 5 objects of complex BRDFs taken from 20 views (100 data in total). Here, we evaluated our method on DiLiGenT-MV with CNN-PS [13] and PS-FCN+ [7] and SPS-Net (32×32 or 8×8) [20] which were trained on our training dataset for the fair comparison [7].The result is illustrated in Fig. 4 and Fig. 6. Because the number of data is 100 (5 objects $\times$ 20 views), we averaged MAE over 100 data for the convenience[8]. The result is basically consistent with what we observed in the evaluation on DiLiGenT main dataset. When we observe the qualitative comparison in Fig. 6, our method consistently produced less noisy surface normal maps while preserving the surface details. We want to emphasize that all the algorithms including ours were trained on the same data and same strategy, therefore the difference simply comes from the architecture. This illustrates that how the attention mechanism from the transformer models and our two-branch model contributed to improve the performance in the sparse photometric stereo problem.

Performance on Dense Problem: Though dense photometric stereo problem (*i.e.*, $m \geq 10$) is out-of-scope of this work, we show the result of applying our model trained on $M = 10$ to the test images whose number is larger than 10. The result is illustrated in Fig. 5. We illustrate the mean angular error (in degrees) for both individual objects (*i.e.*, solid line, left scale) and the average of 10 DiLiGenT objects (*i.e.*, dash-line, right scale). In summary, we didn't observe the significant drop of the prediction accuracy when we input test images whose number is very different from one of training images. However, we didn't observe the significant improvement as well though much larger information is available. In reality, this result is not surprising and coincides with the observation in existing works [7, 29] that models trained on the sparse input set are hard to be generalized to the dense setup. As described in the main paper, we need to increase the number of *training* images to represent more complex interactions among large number of input images. However, the transformer model is well known to be inefficient especially when the number of elements is large, so the adaptation of our model to the dense problem should be left for the future work.

5 Conclusion

In this paper, we presented the self-attention-based photometric stereo network namely PS-Transformer. By incorporating the attention mechanism to capture the complex high-order interactions into the dual-branch designed network to capture both local and global information, our model significantly outperformed any existing deep photometric stereo algorithms in the sparse photometric stereo problem. The current limitation is that the self-attention requires $O(M^2)$ computation therefore our method doesn't scale to the very dense photometric stereo problem. However, developing efficient transformer models is the very hot topic and we believe our PS-Transformer models can benefit from it.

[7] We didn't include GPS-Net in this experiment because the authors implementation was optimized to DiLiGenT dataset and wasn't available on DiLiGenT-MV.

[8] The prediction errors of individual objects and views as well as full qualitative comparison are provided in the supplementary material.

References

[1] N. Alldrin, S. Mallick, and D. Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. *Proc. CVPR*, 2007.

[2] N. Alldrin, T. Zickler, and D. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *Proc. CVPR*, 2008.

[3] B. Burley. Physically-based shading at disney, part of practical physically based shading in film and game production. *SIGGRAPH 2012 Course Notes*, 2012.

[4] J. Chang, J. Gu, L. Wang, G. Meng, S. Xiang, and C. Pan. Structure-aware convolutional neural networks. 2018.

[5] G. Chen, K. Han, and K-Y. K. Wong. Ps-fcn: A flexible learning framework for photometric stereo. *Proc. ECCV*, 2018.

[6] G. Chen, K. Han, B. Shi, Y. Matsushita, and K. K. K. Wong. Self-calibrating deep photometric stereo networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8731–8739, 2019.

[7] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K. Wong. Deep photometric stereo for non-Lambertian surfaces. *TPAMI*, 2020.

[8] CNNPS. https://github.com/satoshi-ikehata/CNN-PS.

[9] D. Goldman, B. Curless, A. Hertzmann, and S. Seitz. Shape and spatially-varying brdfs from photometric stereo. In *Proc. ICCV*, October 2005.

[10] GPS-Net. https://github.com/ZhuokunYao/GPS_NET.

[11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[12] Z. Hui and A. C. Sankaranarayanan. Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(10):2060–2073, 2017.

[13] S. Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proc. ECCV*, 2018.

[14] S. Ikehata and K. Aizawa. Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *Proc. CVPR*, 2014.

[15] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Robust photometric stereo using sparse regression. In *Proc. CVPR*, 2012.

[16] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Photometric stereo using sparse bayesian regression for general diffuse surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(9):1816–1831, 2014.

[17] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proc. ICML*, pages 3744–3753, 2019.

[18] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita. Learning to minify photometric stereo. In *Proc. CVPR*, 2019.

[19] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020. doi: 10.1109/TIP.2020.2968818.

[20] Huiyu Liu, Yunhui Yan, Kechen Song, and Han Yu. Sps-net: Self-attention photometric stereo network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[21] PS-FCN. https://github.com/guanyingc/PS-FCN.

[22] Y. Quéau, T. Wu, F. Lauze, J. D. Durou, and D. Cremers. A non-convex variational approach to photometric stereo under inaccurate lighting. In *Proc. CVPR*, 2017.

[23] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1078–1091, 2014.

[24] B. Shi, Z. Mo, Z. Wu, D.Duan, S-K. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, page (to appear), 2018.

[25] W. M. Silver. *Determining shape and reflectance using multiple images*. Master's thesis, MIT, 1980.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. NIPS (NeurIPS)*, 2017.

[27] P. Woodham. Photometric method for determining surface orientation from multiple images. *Opt. Engg*, 19(1):139–144, 1980.

[28] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Proc. ACCV*, 2010.

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

[30] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pages 3391–3401, 2017.

[31] R. Zhang, P. Tsai, J. Cryer, and M. Shah. Shape-from-shading: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):690–706, 1999.

[32] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L-Y. Duan, and A.C. Kot. Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. *Proc. ICCV*, 2019.