
Reflectance-Guided Deep Photometric Stereo

Ruogu Li
ruoguli

Weison Huang
weisonh

Jiaheng Miao
jiahengm

Joseph Wan
zhilianw

Abstract

Photometric stereo aims to reconstruct an object’s surface normals from multiple images captured under varying lighting directions. Traditional methods often rely on simplistic assumptions about material reflectance properties, limiting their applicability to real-world materials. Recent deep learning approaches [9; 5; 2; 12] have significantly advanced performance over traditional techniques by predicting normal maps directly from input images taken under different lighting conditions. However, these methods typically do not account for material reflectance properties, such as the Bidirectional Reflectance Distribution Function (BRDF), as input. This omission poses an ill-posed problem, as identical image observations can arise from different combinations of surface normals and reflectance properties, hindering generalization across diverse materials. In this work, we propose an extension to existing deep learning photometric stereo methods by explicitly utilizing BRDF as an input. We introduce a feature modulation layer that leverages reflectance information to enhance shape reconstruction accuracy while preserving generalization across a wide range of materials without requiring extensive fine-tuning. *Our approach is designed to seamlessly integrate into any prior deep learning photometric stereo framework*, effectively bridging the gap between traditional material-dependent techniques and modern generalizable deep learning methods. Results show our method achieves up to an 18.7% reduction in mean angular error compared to baseline methods, demonstrating enhanced shape reconstruction accuracy. This offers a robust, efficient, and scalable solution for high-fidelity surface reconstruction. The source code can be found at GitHub repository.¹

1 Introduction

Accurate 3D shape reconstruction is a cornerstone challenge in the fields of computer vision, computer graphics, and robotics. Among the various techniques developed to address this problem, photometric stereo stands out as a powerful method that leverages multiple 2D images captured under varying lighting conditions to estimate an object’s surface normals. Traditional photometric stereo approaches [11] typically rely on simplified analytical models of material reflectance properties. These methods often assume Lambertian reflectance—a model in which light scatters uniformly in all directions—and solve a system of equations to derive surface normals. While effective under controlled conditions, such assumptions limit the applicability of these techniques to real-world surfaces exhibiting complex and diverse reflectance behaviors.

In recent years, deep learning-based methods [9; 5; 2; 12] have emerged as a promising alternative for photometric stereo. These approaches typically take as input a set of images captured under different lighting directions, along with the corresponding lighting vectors, and predict per-pixel normal maps. Notably, these models often do not explicitly account for the reflectance properties of the material, rendering the problem ill-posed. This is because distinct materials may respond differently to identical lighting conditions; for instance, two surfaces with differing normals could produce similar images due to variations in their reflectance characteristics. We hypothesize that this

¹https://github.com/leolrg/idl_project

omission significantly hampers the performance of deep learning models across diverse materials. Moreover, adapting such models to specific materials through fine-tuning is both computationally expensive and data-intensive. Acquiring the necessary training data often requires costly equipment, such as laser scanners [6], further complicating the process.

To address these limitations, we propose a novel approach that explicitly incorporates material reflectance properties into the model. Specifically, we introduce a Feature-wise Linear Modulation (FiLM) and a Conditional Batch Normalization (CBN) layer that transforms the output of convolutional layers differently based on different material’s reflectance properties. This design enables the network to adapt its behavior dynamically to the material characteristics of the object being reconstructed.

The reflectance properties of a material are typically characterized by the Bidirectional Reflectance Distribution Function (BRDF), which defines the ratio of reflected radiance to incident irradiance as a function of incoming and outgoing light directions at each surface point. However, the general form of a BRDF (which is an arbitrary function) is too complex to be directly integrated into a neural network. To overcome this challenge, we adopt the approach proposed in [13], which encodes the BRDF as a 14-dimensional latent vector. This compact representation serves as an effective input to our FiLM/CBN layer, facilitating the integration of reflectance information into the photometric stereo pipeline.

To summarize, we propose a novel network architecture for photometric stereo that directly incorporates BRDF as input, using a 14-dimensional latent representation from [13] fed into a FiLM/CBN layer. *Our results show that explicitly conditioning the network on material reflectance properties significantly improves surface normal estimation, achieving up to an 18.7% reduction in mean angular error compared to the baseline model.* This highlights the importance of integrating material information for accurate and generalizable deep photometric stereo.

2 Related Work and Background

2.1 Mathematical Formulation

Our model for photometric stereo can be formulated as learning a mapping function:

$$f : \text{Agg}(\mathbf{M}^i, \mathbf{l}^i, \Phi) \rightarrow \mathbf{N}, i \in \{1, 2, \dots, n\} \quad (1)$$

where $\mathbf{M}^i \in \mathbb{R}^{h \times w \times 3}$ represents the observed RGB image under lighting direction $\mathbf{l}^i \in \mathbb{R}^3$ (represented as a unit vector), $\Phi \in \mathbb{R}^{14}$ represents a parameterization of the material’s BRDF, and $\mathbf{N} \in \mathbb{R}^{h \times w \times 3}$ represents the per-pixel surface normal. The Agg function aggregates the information from multiple observations. Previous deep learning methods for photometric stereo do not incorporate the Φ parameter, thus failing to explicitly account for the reflectance properties of materials.

In comparison, the traditional approach [11] is based on the physical image formation model. Let’s consider a scene under orthographic projection with directional lighting and camera-facing viewing direction. For non-Lambertian surfaces, the image formation process for a pixel (i, j) observed under light direction $\mathbf{l}$ can be modeled as:

$$m_{i,j} = \rho(\mathbf{n}_{i,j}, \mathbf{l}, \mathbf{v}) \cdot \max\{\mathbf{n}_{i,j}^\top \mathbf{l}, 0\} + \epsilon \quad (2)$$

where ρ represents the BRDF that characterizes how light interacts with the surface (see 2.3 for a detailed explanation of BRDF), $\mathbf{n}_{i,j}$ is the surface normal at pixel (i, j) , $\mathbf{v}$ is the viewing direction, and ϵ accounts for global illumination effects and noise. The $\max\{\mathbf{n}_{i,j}^\top \mathbf{l}, 0\}$ term is for modeling attached shadows. Using at least three input images and a fixed BRDF model, this traditional method solves the image model (i.e. the above equation) inversely in order to compute the surface normals.

By explicitly incorporating BRDF information in our learning framework, our method bridges the gap between material-specific and general-purpose models. This approach allows our model to directly learn how reflectance influences the observed intensities, in addition to learning the influence of surface geometry.

2.2 Deep Learning Method

The field of photometric stereo was first introduced by Woodham [11] in 1980, who demonstrated that surface normals can be recovered from images taken under varying lighting conditions with a fixed

viewpoint. However, Woodham’s method was constrained by the Lambertian reflectance assumption, which limited its practical applications. The field remained constrained by these limitations until DPSN [9] first attempted to leverage deep learning for the task of photometric stereo. However, DPSN [9] required a predefined set of fixed lighting directions and could only process a specific number of input images due to the design of its architecture, limiting its practical application. In 2018, Ikehata addressed some of DPSN’s shortcomings with CNN-PS [5], a per-pixel approach that incorporated confidence estimation to handle challenging surface properties like self-shadows and specularities. While more robust than DPSN [9], CNN-PS [5] still processed each pixel in isolation without considering the spatial context.

A parallel breakthrough came with PS-FCN [2] by Chen et al. in 2018, which introduced an innovative all-pixel approach that processes entire images holistically using CNN rather than inspecting individual pixels. PS-FCN’s [2] key innovation was its max-pooling fusion layer, which enables the network to handle an arbitrary number of input images in an order-agnostic manner. By taking the maximum value across feature maps at pixel position, the network is able to handle an arbitrary number of input images in an order-agnostic manner because the max operation is invariant to both the number and sequence of inputs. PS-FCN’s [2] architectural design offers significantly improved flexibility compared to DPSN [9], which makes PS-FCN [2] very practical. Moreover, PS-FCN’s [2] all-pixel approach takes the spatial context of surrounding pixels into account.

Despite these advances, current deep learning models such as PS-FCN [2] do not explicitly incorporate reflectance properties as input. As a result, the current models often perform poorly on materials with different reflectance properties than those in the training set. We aim to extend PS-FCN [2] by directly integrating BRDF information into the network architecture. To evaluate the effectiveness of our proposed approach, we select the original PS-FCN architecture as our primary baseline, which serves as a reference to assess the improvement brought by our integration of reflectance properties.

2.3 BRDF Representation

The Bidirectional Reflectance Distribution Function (BRDF) is a widely used formalism for characterizing the reflectance properties of materials. Defined as a function $f(\omega_i, \omega_r)$, the BRDF quantifies the ratio of reflected radiance along a viewing direction ω_r to the incident radiance from an incoming light direction ω_i . Traditional photometric stereo methods often model the BRDF using simple analytical functions. However, these models typically fail to capture the intricate reflectance behavior exhibited by real-world materials. To address this limitation, empirical datasets have been developed to measure BRDFs of physical materials. For instance, the MERL dataset [7] provides densely sampled BRDFs for 100 isotropic real-world materials, capturing reflectance across a wide range of incoming and viewing directions. Similarly, the EPFL dataset [4] includes BRDF measurements for 62 isotropic and anisotropic materials.

Measured BRDFs are typically stored as large, discrete, high-dimensional matrices or tensors, with variations in density and structure depending on the sampling methodology. This representation poses challenges for direct integration into neural networks due to its size and complexity. To mitigate this, several approaches have been proposed to compress measured BRDFs into more manageable forms. Notable methods, such as those in [1; 10], employ Principal Component Analysis (PCA) to derive compact representations from the high-dimensional BRDF matrices.

In this work, we focus on the approach introduced in [13], which leverages a neural network to encode measured BRDFs of arbitrary size and density into a fixed-length latent vector. Specifically, given a measured BRDF represented as a set of observations $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_n\}$, where $x_k \in \mathbb{R}^4$ encodes the k -th pair of incoming and viewing directions and $y_k \in \mathbb{R}$ denotes the corresponding ground-truth reflected radiance ratio, the method constructs an encoder through two key components. First, a shared-weight multilayer perceptron (MLP), denoted h , processes each observation (x_k, y_k) to map it into a 14-dimensional feature vector, producing the set $\{h(x_1, y_1), \dots, h(x_n, y_n)\}$. Second, an aggregation function a , implemented as an element-wise mean across the n feature vectors, combines these outputs into a single fixed-size latent vector $z = a(h(x_1, y_1), \dots, h(x_n, y_n)) \in \mathbb{R}^{14}$. Together, h transforms individual observations into a consistent feature space, while a ensures invariance to the number of input observations by summarizing them into a unified representation, thus forming a robust encoder capable of handling BRDFs of varying sizes. A decoder, implemented as another MLP, takes the latent vector z along with a direction pair (x_k) as input and predicts the corresponding radiance ratio $\hat{y}_k$. The network is trained end-to-end

by minimizing the difference between the predicted radiance ratios $\hat{y}_k$ and the ground-truth values y_k across all k observations, typically using a loss function such as mean squared error. Training is performed on both the MERL [7] and EPFL [4] datasets, enabling the encoder to generalize across diverse material types. Once trained, the encoder can generate a latent vector z from any arbitrarily sized measured BRDF, providing a versatile and compact representation for downstream tasks.

3 Methodology

3.1 Data

3.1.1 Training Dataset: PS-FCN Blobby Shapes with MERL BRDF

We train our model on the public PS-FCN Blobby Shape dataset rendered with real-world materials from the MERL BRDF collection [7]. This dataset comprises 10 distinct “blobby” meshes, each observed from a dense grid of viewpoints: a total of 1,296 views per shape (36 azimuth angles $\times$ 36 elevation angles). For each view, two BRDFs are randomly sampled (out of 100), yielding 25,920 unique samples (10 shapes $\times$ 1,296 views $\times$ 2 BRDFs). Each sample consists of 64 images at 128 $\times$ 128 px resolution rendered under light directions uniformly sampled over a 180 $^\circ$ $\times$ 180 $^\circ$ hemisphere.

3.1.2 Evaluation Dataset A: PS-FCN Sculpture Shapes with MERL BRDF

To assess generalization to more complex geometries, we evaluate on the PS-FCN Sculpture dataset [10], rendered under the same MERL BRDFs and lighting regime as above. This set contains 8 complicated sculpture models. Each full-shape render is produced at 512 $\times$ 512 px, which is then extracted to 128 $\times$ 128 px patches (discarding those with $< 50\%$ foreground) to form 59,292 valid samples, each with 64 images and corresponding normal maps and lighting directions. We use this entire set for testing.

3.1.3 Evaluation Dataset B: Blobby Shapes with EPFL BRDF

To evaluate robustness to novel unseen material distributions, we synthesized an additional test set using the EPFL BRDF dataset [4]. Starting from the same 10 PS-FCN blobby meshes, we render 49 viewpoints per shape (7 azimuth $\times$ 7 elevation) for 490 renders. For each render, one BRDF is randomly selected from EPFL and images are produced under 32 random light directions at 128 $\times$ 128 px. This yields 490 $\times$ 32 = 15,680 images, each paired with its BRDF identifier, lighting direction, and ground-truth normal map.

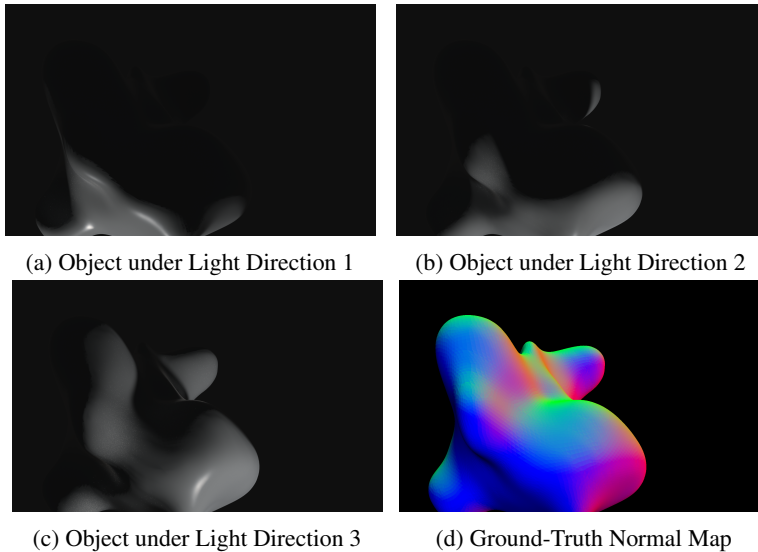


Figure 1: Example in EPFL synthetic dataset. (a), (b), (c) is a sample generated from 3 different lighting directions and (d) is the ground truth normal map

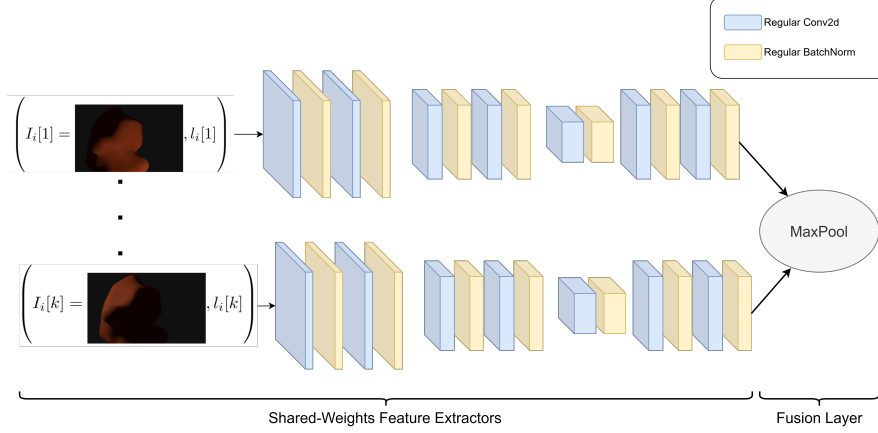


Figure 2: Feature Extractors and Fusion Layer for Baseline Model: PS-FCN

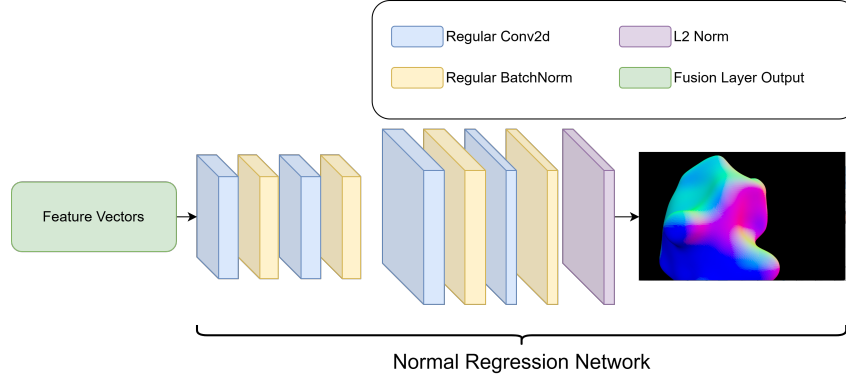


Figure 3: Normal Regression Network for Baseline PS-FCN AND Our Model

3.2 Baseline Model Architecture: PS-FCN

To assess the effectiveness of integrating BRDF via Conditional Batch Normalization (CBN), we select the original PS-FCN network [2] as our primary baseline. As displayed in Figure 2 and Figure 3, the original PS-FCN comprises a shared-weight feature extractor, a max-pooling fusion layer, and a normal regression network. This architecture design renders PS-FCN flexible in accepting an arbitrary number of images captured under various light directions. Specifically, for an object captured under k light directions (with k arbitrarily chosen), each light direction (a 3-vector) is broadcasted and concatenated with its corresponding image. This operation yields an input tensor of dimensions $k \times 6 \times H \times W$, which effectively integrates the appearance information from images with the illumination data.

The PS-FCN network first passes the input tensor into a shared-weight convolutional network to extract feature maps from each image–light pair independently. This feature extractor is composed of several convolutional layers, interleaved with downsampling and upsampling operations, to capture both local and global contextual information critical for photometric stereo. Importantly, the same network is used for each input image–light pair.

After the Shared-Weights Feature Extractor, the network aggregates the feature maps from all image–light pairs using a channel-wise max-pooling operation (referred to as Fusion Layer in Figure 2). This fusion strategy is particularly effective in photometric stereo because it does not rely on a specific number of inputs, allowing the network to handle a variable number of image–light pairs. By selecting the maximum activation across channels at each spatial location, the Fusion Layer consolidates the most informative features from multiple observations into a unified representation, forming a solid foundation for the subsequent the network to generate high-quality surface normal maps.

The fused feature map is then processed by a normal regression network (as displayed in Figure 3), which is a series of convolutional layers that upsample and refine the features to produce the predicted normal map of the object.

3.3 Proposed Model 1: PS-FCN with Conditional BatchNorm

We propose a network architecture that extends the PS-FCN framework by integrating material-specific reflectance information, expressed as a BRDF feature vector, directly into the network. To achieve this, we adopted the method of Conditional Batch Normalization (CBN) [3], which modulates the network’s feature maps using external conditioning. This allows our model to adapt its internal representations based on the reflectance properties of the material, thus improving the accuracy of reconstruction.

Our proposed model is similar to the original PS-FCN, but with additional conditional batch normalization layers in the feature extractor to utilize BRDF information, as shown in (Figure 4a, 3). In contrast to the original PS-FCN, which accepts only image-light pairs as input, our model takes in a 14-dimensional BRDF latent vector (which is obtained as described in Section 2.3) as an additional input. The BRDF vector is processed by a multi-layer perceptron (MLP) to produce an embedding (call it e_{BRDF}) that will be used to modulate the network at each Conditional BatchNorm layer, as labeled in Figure 4a.

To integrate the material-specific information, we replace the conventional Batch Normalization (BN) layers within the feature extractor with the Conditional Batch Normalization layers (CBN) [3], which operate based on the BRDF embedding e_{BRDF} , as shown in Figure 4b. Recall that given a mini-batch of feature maps $B = \{F_{i,c,h,w}\}_{i=1}^N$, standard BN normalize them as follows:

$$\text{BN}(F_{i,c,h,w} \mid \gamma_c, \beta_c) = \gamma_c \frac{F_{i,c,h,w} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta_c,$$

where γ_c and β_c are learned scaling and shifting parameters. In our approach, the BRDF embedding vector e_{BRDF} is first fed through two MLPs that predict small adjustments $\Delta\gamma$ and $\Delta\beta$ (Figure 4b), respectively, for each Conditional BatchNorm Layer. These adjustments are then added to the original BN parameters:

$$\hat{\gamma}_c = \gamma_c + \Delta\gamma_c, \hat{\beta}_c = \beta_c + \Delta\beta_c.$$

The resulting Conditional Batch Normalization (CBN) operation then becomes:

$$\text{BN}(F_{i,c,h,w} \mid \hat{\gamma}_c, \hat{\beta}_c).$$

By integrating CBN at multiple layers within the feature extractor, the network is able to modulate both low-level and high-level features based on the material’s reflectance properties. This approach is lightweight, and while we have applied it specifically to PS-FCN, it can be seamlessly extended to any photometric stereo network.

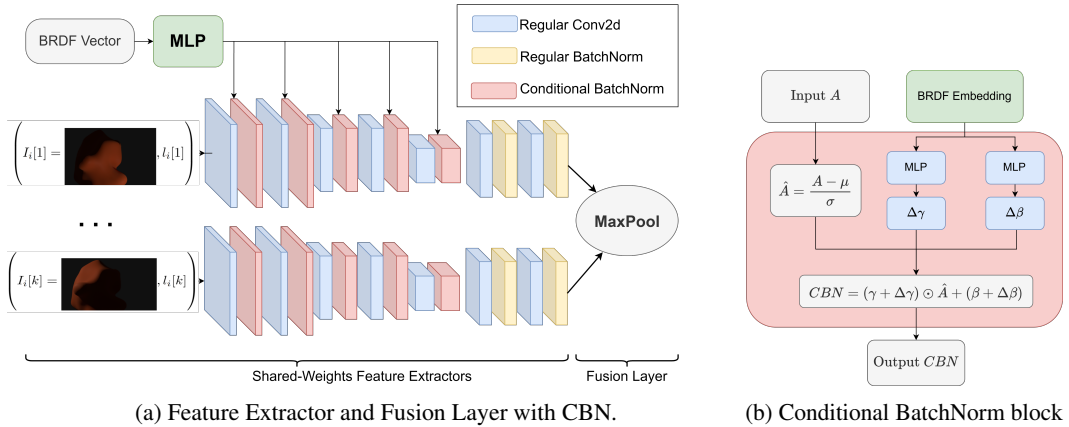


Figure 4: Model architecture: PS-FCN with Conditional BatchNorm.

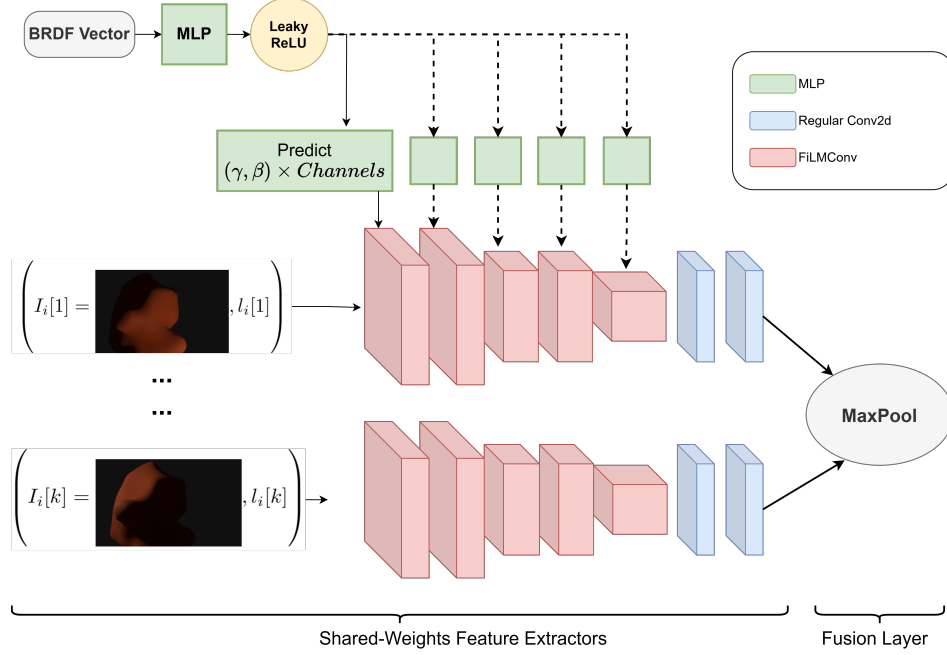


Figure 5: Feature Extractors and Fusion Layer incorporating Feature-wise Linear Modulation

3.4 Proposed Model 2: PS-FCN with Feature-wise Linear Modulation

We propose an alternative feature-modulation mechanism for PS-FCN by replacing Conditional Batch Normalization with Feature-wise Linear Modulation (FiLM) [8], a general conditioning layer that applies an affine transform to each feature channel based on an external input. As shown in Figure 5, the 14-dimensional BRDF vector is first passed through an MLP with the Leaky ReLU activation function to produce a latent embedding e_{BRDF} , which is then used to condition each convolutional block via FiLM.

Formally, let $F^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$ denote the output of the l -th convolution block within the Shared-Weights Feature Extractors of the PS-FCN model. FiLM learns two functions $f^{(l)}$ and $h^{(l)}$ which output $\gamma^{(l)}$ and $\beta^{(l)}$, where $(\gamma^{(l)}, \beta^{(l)}) \in \mathbb{R}^{C_l}$ as a function of input e_{BRDF} , respectively:

$$\gamma^{(l)} = f^{(l)}(e_{\text{BRDF}}), \quad (3)$$

$$\beta^{(l)} = h^{(l)}(e_{\text{BRDF}}), \quad (4)$$

where both $f^{(l)}, h^{(l)}$ are implemented as MLPs. These parameters then modulate the activations via a feature-wise affine transform:

$$\text{FiLM}(F_{c,h,w}^{(l)} \mid \gamma_c^{(l)}, \beta_c^{(l)}) = \gamma_c^{(l)} F_{c,h,w}^{(l)} + \beta_c^{(l)} \quad (5)$$

where $c = 1, \dots, C_l$, $h = 1, \dots, H_l$, $w = 1, \dots, W_l$. In practice, it is more convenient to merge $f^{(l)}$ and $h^{(l)}$ into a single function that takes in e_{BRDF} and outputs one vector $(\gamma^{(l)}, h^{(l)})$. For the rest of the model architecture, both the fusion (max-pooling) and normal-regression head remain the same as in the original PS-FCN network.

As FiLM provides explicit, per-layer affine control, we expect that it can adapt to diverse, non-Lambertian reflectance behaviors more flexibly compared with CBN, which conditions on batch statistics and can inadvertently wash out sample-specific shading cues when batch composition varies. A quantitative ablation comparing FiLM and CBN variants will be displayed in Section 5.

3.5 Evaluation Metric

We evaluate our method using the mean angular error (MAE) measured in degrees. MAE quantifies the average angular deviation between the predicted and ground truth normal maps. Since the

synthetic dataset includes irrelevant background regions (i.e., black areas in Figure 1), we apply a mask to focus exclusively on the foreground regions.

Given the predicted normal map $\tilde{\mathbf{N}}$ and ground truth normal map $\mathbf{N}$, MAE is computed as:

$$\text{MAE} = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \arccos \left(\frac{\mathbf{N}_{i,j} \cdot \tilde{\mathbf{N}}_{i,j}}{|\mathbf{N}_{i,j}| |\tilde{\mathbf{N}}_{i,j}|} \right) \times \frac{180^\circ}{\pi} \quad (6)$$

where Ω represents the foreground pixel set and $\arccos(\cdot)$ computes the angle between the normalized vectors in radians, which is then converted to degrees. Lower MAE indicates better performance, directly measuring the accuracy of surface normal reconstruction.

3.6 Loss Function

We use cosine embedding loss as our loss function. Specifically, given the ground-truth normal map $\mathbf{N}$ and predicted normal map $\tilde{\mathbf{N}}$, the cosine embedding loss is defined as

$$L_{\text{normal}} = \frac{1}{hw} \sum_{i,j} \left(1 - \mathbf{N}_{i,j} \cdot \tilde{\mathbf{N}}_{i,j} \right), \quad (7)$$

where $\mathbf{N}_{i,j}$ and $\tilde{\mathbf{N}}_{i,j}$ represents the ground-truth and predicted normal map at pixel i, j . If the predicted normal closely aligns with the ground truth, their dot product $\mathbf{N}_{i,j} \cdot \tilde{\mathbf{N}}_{i,j}$ approaches 1, leading to a lower loss. Conversely, if $\mathbf{N}_{i,j}$ and $\tilde{\mathbf{N}}_{i,j}$ are completely opposite (i.e a difference of 180 degrees), their dot product is -1 , leading to a large loss.

4 Experiments

4.1 Training Method

We trained our proposed models, PS-FCN with CBN and PS-FCN with FiLM, on the PS-FCN Blobby Shapes dataset rendered with MERL BRDFs (described in Section 3.1). To leverage prior knowledge and accelerate convergence, the convolutional layers in the shared feature extractor and the normal regression network were initialized using weights from a publicly available PS-FCN model pre-trained on the same dataset (without BRDF conditioning). The newly introduced MLP layers used for generating the conditioning parameters ($\Delta\gamma, \Delta\beta$ for CBN; γ, β for FiLM) were initialized randomly.

Training was performed using the Adam optimizer with a batch size of 16. The Cosine Embedding Loss (detailed in Section 3.6) was used as the objective function.

For the PS-FCN with FiLM model, training proceeded for 50 epochs. The initial learning rate was set to 1×10^{-4} . We employed a learning rate scheduler that reduced the learning rate by a factor of 0.5 every 5 epochs.

For the PS-FCN with CBN model, training was conducted for 30 epochs. The initial learning rate was 1×10^{-3} , and it was reduced by a factor of 0.1 every 10 epochs.

5 Results and Analysis

In this section, we present a comprehensive evaluation of our approach, demonstrating how integrating BRDF information through feature modulation techniques enhances photometric stereo performance.

5.1 Quantitative Results

Table 1 presents the mean angular error (in degrees) for each model across the three datasets. Lower values indicate better performance.

Table 1: Mean Angular Error (in degrees) on different datasets

Model	Blobby (MERL)	Sculpture (MERL)	Blobby (EPFL)
PS-FCN (Baseline)	5.25	9.63	13.56
PS-FCN with CBN (Ours)	4.89	9.14	14.80
PS-FCN with FiLM (Ours)	4.27	8.51	14.01

Our results demonstrate that both feature modulation approaches (CBN and FiLM) outperform the baseline PS-FCN on datasets with materials seen during training (Blobby MERL and Sculpture MERL). The PS-FCN with FiLM consistently achieves the best performance, reducing the mean angular error by 18.7% on the Blobby (MERL) dataset and 11.6% on the Sculpture (MERL) dataset compared to the baseline. However, both our proposed models underperform the baseline on the Blobby (EPFL) dataset, which contains novel materials not seen during training. This suggests that while our approach excels at utilizing reflectance properties for materials in the training set, it struggles to generalize to entirely new materials. This limitation likely stems from the sparsity of our training BRDF samples (100 materials) in the 14-dimensional latent space, making interpolation to unseen materials challenging.

Our experiments show that FiLM consistently outperforms CBN across datasets. We attribute this to FiLM’s more direct approach to feature modulation, which directly scales and shifts feature maps based on BRDF information, compared to CBN’s indirect approach of modulating batch normalization parameters.

5.2 Training and Validation Performance

Figure 6 illustrates the training and validation error as well as the learning rates of the CBN-based model and the FiLM-based model over the course of training.

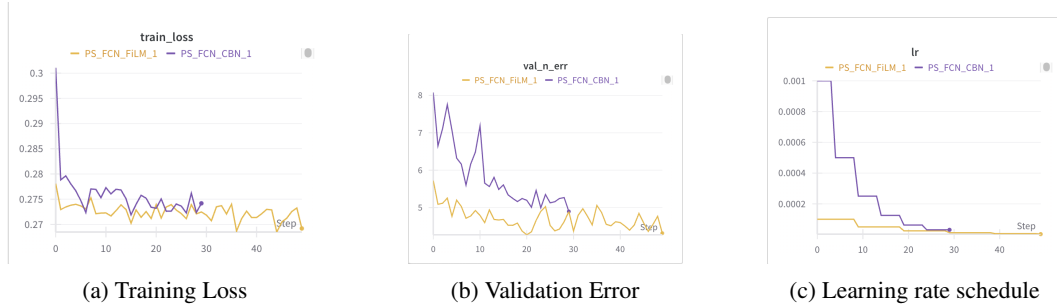


Figure 6: Training logs

5.3 Significance of Results

The improvements demonstrated by our approach address a fundamental limitation in current deep learning-based photometric stereo methods. By explicitly accounting for material properties, we transform what was previously an ill-posed problem into one that is better constrained. The 18.7% reduction in mean angular error on the Blobby (MERL) dataset and 11.6% reduction on the Sculpture (MERL) dataset highlight the effectiveness of our approach in practical scenarios. Our work effectively bridges the gap between classical photometric stereo methods, which explicitly model reflectance properties but struggle with complex materials, and modern deep learning approaches, which handle complex materials but ignore explicit reflectance models. By combining the strengths of both paradigms, we achieve more robust performance.

The feature modulation approach we propose can be seamlessly integrated into various existing deep learning frameworks for photometric stereo without requiring fundamental architectural changes. This modularity makes our contribution highly practical and adaptable. Moreover, the consistent outperformance of FiLM over CBN across our experiments provides valuable insight into the most

effective mechanisms for integrating auxiliary information into deep neural networks. This finding may inform future work in conditional computation beyond photometric stereo.

5.4 Potential Risks and Limitations

Despite the promising results, our approach has a limitation that should be acknowledged. Our models’ underperformance on the Blobby (EPFL) dataset reveals a significant limitation—the inability to generalize effectively to materials not seen during training. This suggests that our approach may require retraining or fine-tuning when applied to scenes with novel material properties. We believe that this limitation can be resolved by training the model with more materials, which improves generalization.

6 Future Directions

Our work demonstrates the significant benefit of incorporating explicit material reflectance information (BRDFs) into deep photometric stereo networks using feature modulation techniques like CBN and FiLM. The results show marked improvements in accuracy when evaluated on materials similar to those seen during training. However, the observed performance drop on the EPFL dataset highlights a key limitation: generalizing to fundamentally novel, unseen BRDFs remains a challenge. Based on these findings, we identify several promising avenues for future research:

1. **Enhancing Generalization to Novel Materials:** The primary limitation is the model’s struggle with out-of-distribution BRDFs. This likely stems from the relatively sparse sampling of the vast material space provided by the 100 MERL BRDFs used for training. Future work should focus on:
 - **Training with Denser Material Representations:** Augmenting the training data with a much larger and more diverse set of BRDFs is crucial. This could involve incorporating additional measured datasets or, more scalably, leveraging analytical BRDF models (e.g., microfacet models like Cook-Torrance) to synthetically generate a dense distribution of plausible BRDFs by varying their physical parameters (roughness, metallicness, index of refraction, etc.). This would provide the network with broader coverage of the material space, potentially improving its ability to interpolate or extrapolate to unseen materials.
 - **Improving BRDF Latent Representation:** Investigating alternative or refined BRDF encoding methods beyond the current 14-dimensional latent vector could yield representations that capture material properties more effectively for generalization. This might involve exploring different encoder architectures, latent dimensionalities, or loss functions during the BRDF encoding phase, specifically optimized for downstream photometric stereo tasks. Hybrid representations combining latent vectors with key analytical parameters might also be beneficial.
2. **Exploring Advanced Conditioning Mechanisms:** While FiLM outperformed CBN, further research could explore more sophisticated conditioning mechanisms. Techniques like attention-based modulation, where the network learns to dynamically weigh the influence of different BRDF features based on the input image content, or adaptive instance normalization variants tailored for BRDF conditioning, might offer more nuanced control and better adaptation.

7 Conclusion

Our work successfully addressed the fundamental limitation in deep learning-based photometric stereo by explicitly integrating material reflectance properties through feature modulation. Our approach achieved up to 18.7% reduction in mean angular error compared to baseline methods, demonstrating that accounting for BRDF information significantly enhances surface normal reconstruction accuracy. The proposed architecture offers a modular solution that can be seamlessly integrated into existing frameworks, effectively transforming what was previously an ill-posed problem into one that is better constrained. While challenges remain in generalizing to entirely novel materials, our work establishes a promising direction for unifying physics-based material understanding with deep learning techniques for high-fidelity surface reconstruction.

References

- [1] Mahdi M Bagher, John Snyder, and Derek Nowrouzezahrai. A non-parametric factor microfacet model for isotropic brdfs. *ACM Transactions on Graphics (TOG)*, 35(5):1–16, 2016.
- [2] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–18, 2018.
- [3] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. *Advances in neural information processing systems*, 30, 2017.
- [4] Jonathan Dupuy and Wenzel Jakob. An adaptive parameterization for efficient material acquisition and rendering. *ACM Transactions on graphics (TOG)*, 37(6):1–14, 2018.
- [5] Satoshi Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–18, 2018.
- [6] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials, 2020.
- [7] Wojciech Matusik. *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003.
- [8] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [9] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo network. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 501–509, 2017.
- [10] Tiancheng Sun, Henrik Wann Jensen, and Ravi Ramamoorthi. Connecting measured brdfs to analytic brdfs by data-driven diffuse-specular separation. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.
- [11] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.
- [12] Ru Yang, Yaoke Wang, Shuheng Liao, and Ping Guo. Dpps: A deep-learning based point-light photometric stereo method for 3d reconstruction of metallic surfaces. *Measurement*, 210:112543, 2023.
- [13] Chuankun Zheng, Ruzhang Zheng, Rui Wang, Shuang Zhao, and Hujun Bao. A compact representation of measured brdfs using neural processes. *ACM Transactions on Graphics (TOG)*, 41(2):1–15, 2021.

Appendix

Team Contributions:

Ruogu Li is the lead researcher of this project. Weison Huang, Jiaheng Miao, and Joseph Wan are research assistants, providing experimental and writing support.

Github Repository:

https://github.com/leolrg/idl_project