Learning Efficient Photometric Feature Transform for Multi-view Stereo

Kaizhang Kang[1], Cihui Xie[1], Ruisheng Zhu[1], Xiaohe Ma[1], Ping Tan[3], Hongzhi Wu[1*] and Kun Zhou[1,2]

[1]Zhejiang University   [2]ZJU-FaceUnity Joint Lab of Intelligent Graphics   [3]Simon Fraser University

Abstract

We present a novel framework to learn to convert the per-pixel photometric information at each view into spatially distinctive and view-invariant low-level features, which can be plugged into existing multi-view stereo pipeline for enhanced 3D reconstruction. Both the illumination conditions during acquisition and the subsequent per-pixel feature transform can be jointly optimized in a differentiable fashion. Our framework automatically adapts to and makes efficient use of the geometric information available in different forms of input data. High-quality 3D reconstructions of a variety of challenging objects are demonstrated on the data captured with an illumination multiplexing device, as well as a point light. Our results compare favorably with state-of-the-art techniques.

1. Introduction

As one central problem in computer vision and graphics, shape reconstruction in the presence of complex appearance is challenging. At one hand, multi-view stereo methods [8] usually require a Lambertian-dominant reflectance for computing reliable view-invariant features. The appearance variation with view or lighting is undesirable, as it may change the native spatial features on the object, or specularly reflect the projected pattern from active illumination [16, 23], leading to difficulties in correspondence computation for shape reconstruction.

On the other hand, single-view photometric stereo [34, 5] exploits the lighting variations on each pixel and transforms the image measurements into a normal map. While high-quality details can be recovered, it suffers from low-frequency shape distortions [20]. Recently, multi-view photometric stereo [12, 32, 17] accurately integrates the photometric cues from different viewpoints, subject to the geometric constraints across multiple views.

However, photometric stereo techniques are not scalable

to the amount of geometric information in the input data for 3D reconstruction, leading to suboptimal results. When additional physical cues such as rapid albedo variations are present, they cannot be exploited to improve the reconstruction quality. On the other hand, when the measured photometric information is insufficient to determine a normal field, the quality of results from existing approaches will significantly degrade. Furthermore, related techniques heavily exploit reflectance properties [17, 18], which hinders the extension to handle more general appearance such as anisotropic materials.

To tackle the above challenges, we make the key observation that it is not necessary to use **normal** as the intermediate representation for 3D reconstruction from photometric measurements. Instead, we propose a novel differentiable framework, to efficiently transform the per-pixel photometric information measured at each view, into **automatically learned low-level features**, in an end-to-end fashion. The learned per-pixel features essentially exploit the available geometric information in photometric measurements, and can be plugged in existing multi-view stereo pipelines for further processing tasks like spatial aggregation, resulting in enhanced geometric reconstruction. Furthermore, our data-driven framework is highly flexible and can adapt to various factors, including the physical acquisition capabilities / characteristics of different setups, and different types of appearance such as anisotropic reflectance, by feeding corresponding training data.

The effectiveness of our framework is demonstrated with a high-performance illumination multiplexing device, on geometric reconstruction results of a variety of 3D objects using as few as 16 input photographs per view. Moreover, the framework is generalized to handle the input data of conventional photometric stereo with one point light on at a time (DiLiGenT-MV[17]). Our results compare favorably with state-of-the-art techniques. We make public the code and data of this project [1].

---

*: corresponding author (hwu@acm.org).

[1]https://svbrdf.github.io/publications/ptmvs/project.html

2. Related work

2.1. Multi-view Stereo

These methods [8] first compute low-level features in multi-view input images, which ideally should be discriminative in the spatial domain, and invariant with respect to various factors. Next, the correspondences between features at different views are established, and used to determine 3D points via triangulation. Since the raw measurements at a single pixel are usually not sufficient to accurately establish multi-view correspondences, spatial aggregation is typically performed to incorporate more information from neighboring pixels.

Excellent results with Lambertian-dominant materials are demonstrated from photographs taken with even uncontrolled conditions [9, 24, 25]. But these methods fail on textureless objects, as the features across different locations may be almost identical. Active lighting methods, such as laser-stripe triangulation [16] or structured lighting [23], handle such cases by physically projecting spatially distinctive patterns onto the object. Additional physical dimensions like polarization can also be exploited [6]. Recently, machine learning further pushes the reconstruction quality, by replacing hand-crafted features with automatically learned ones [30, 38, 37, 31]. However, for complex appearance that changes with view / lighting conditions, it is still challenging to compute distinctive and invariant features for 3D reconstruction.

Our work is orthogonal to the majority of work here, which focuses on processing information in the spatial domain. Instead, we focus on the angular domain, by learning to transform the per-pixel photometric information into useful features. We employ one existing pipeline for subsequent spatial-domain processing, and leave the unified treatment of both domains for future work. It is worth mentioning that the work of [36] transforms multi-view images of a homogeneous isotropic reflectance into a diffuse one, to reduce the view variance of specular reflections.

2.2. Photometric Stereo

This class of techniques compute an accurate normal field that can be subsequently integrated into a depth map, from appearance variations under typically a large number of different illumination conditions [28]. Starting from the seminal work of [34] with a Lambertian reflectance and a calibrated directional light, substantial efforts been made to extend to handle more general materials [1, 10, 27] and/or unknown lighting conditions [2, 3, 19].

The closest work to ours is multi-view photometric stereo. While traditional approaches consider a single view only, it combines photometric cues at multiple different views to produce a complete 3D shape. An initial coarse geometry is refined with the normal information, as in [12, 39, 17]. The depth maps integrated from normal fields are directly fused to obtain the final result, according to [32]. And the relationship between a signed distance field and normals is exploited in [18].

Our work also takes as input photometric information for 3D reconstruction. The main difference is that we do not use a fixed intermediate representation (e.g., normal), and instead compute learned low-level features that efficiently scale to the amount of geometric information available in the input data. Moreover, inspired by recent work on differentiable reflectance capture [15], our physical acquisition process can be jointly optimized with the computational transformation of features.

3. Acquisition Setup

Our main experiments are conducted with a high-performance, box-shaped lightstage, similar to [15]. Its size is 80cm × 80cm × 77cm. The sample object is placed on a digital turntable near the center of the device, and rotated to different angles for multi-view imaging. A single FLIR BFS-U3-123S6C-C vision camera captures photographs at a resolution of 4,096×3,000. We illuminate the sample with 24,576 LEDs on the six faces of the device with polycarbonate diffusers attached. The total LED power is about 2,000W, and the pitch of adjacent LEDs is 1cm. We calibrate the intrinsic / extrinsic parameters of the camera, as well as the positions, orientations, angular intensity of LEDs. The rotation angle of the turntable is computed from printed markers on its surface [7]. Please refer to Fig. 1 for an illustration.

4. Preliminaries

The following derivations are based on a gray-scale channel. First, the outgoing radiance B from a surface point $\mathbf{p}$ towards a camera can be modeled as [22]:

$$B(I; \mathbf{p}) = \sum_l I(l) \int \frac{1}{||\mathbf{x_l} - \mathbf{x_p}||^2} \Psi(\mathbf{x_l}, -\omega_\mathbf{i}) V(\mathbf{x_l}, \mathbf{x_p})$$
$$f(\omega_\mathbf{i}'; \omega_\mathbf{o}', \mathbf{p})(\omega_\mathbf{i} \cdot \mathbf{n_p})^+ (-\omega_\mathbf{i} \cdot \mathbf{n_l})^+ d\mathbf{x_l}. \quad (1)$$

Here each light l is a locally planar source. $\mathbf{x_p}$ / $\mathbf{n_p}$ is the position / normal of $\mathbf{p}$, while $\mathbf{x_l}$ / $\mathbf{n_l}$ is the position / normal of a point on the light source l. $\omega_\mathbf{i}$ / $\omega_\mathbf{o}$ denotes the lighting / view direction in the world space, and $\omega_\mathbf{i}'$ / $\omega_\mathbf{o}'$ is expressed in the local frame of $\mathbf{p}$. Note that $\omega_\mathbf{i} = \frac{\mathbf{x_l} - \mathbf{x_p}}{||\mathbf{x_l} - \mathbf{x_p}||}$. $I(l)$ is the intensity for the light l, in the range of [0, 1]. The array $\{I(l)\}$ corresponds to a **lighting pattern**. $\Psi(\mathbf{x_l}, \cdot)$ represents the angular distribution of the light intensity. V is a binary visibility function between $\mathbf{x_l}$ and $\mathbf{x_p}$. The operator $(\cdot)^+$ computes the dot product between two vectors, and clamps any negative result to zero. $f(\cdot; \omega_\mathbf{o}', \mathbf{p})$ is a 2D BRDF slice, which is a function of the lighting direction. In this paper, we employ a standard anisotropic GGX

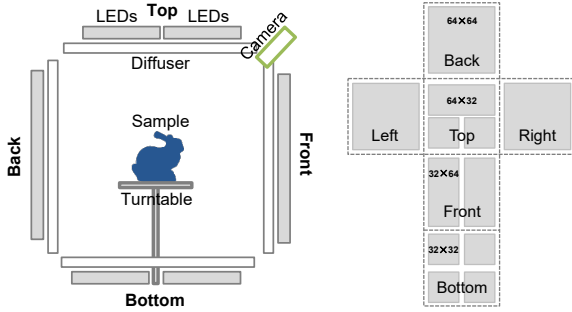model [33] to represent f. Note that our framework is not tied to the choice of the BRDF model.

Figure 1: The lighting layout of our setup. A side view of the setup (left) and the vertical-cross parameterization of all lights with 4,096 LEDs on each face (right).

As B is linear with respect to I (Eq. 1), it can be expressed as the dot product between I and a **lumitexel** c:

$$B(I; \mathbf{p}) = \sum_l I(l)c(l; \mathbf{p}), \qquad (2)$$

where c is a function of the light source l, defined on the surface point $\mathbf{p}$ of the sample object:

$$c(l; \mathbf{p}) = B(\{I(l) = 1, \forall_{j \neq l} I(j) = 0\}; \mathbf{p}). \qquad (3)$$

Each element of c records a measurement B with one light on at a time.

5. Overview

We propose a mixed-domain neural network, to physically encode the photometric information into a small number of measurements by projecting learned lighting patterns, and then computationally transform spatially discriminative and view/lighting-invariant low-level features, on a per-pixel basis. The procedure is repeated for each pixel in each image taken at a particular view, resulting in multi-view, high-dimensional transformed feature maps. These maps are then post-processed and fed into an existing multi-view stereo technique to produce the final 3D geometry. Fig. 2 illustrates the process.

While the network is originally designed to work in conjunction with our lightstage, we will describe how to generalize to handle the input of conventional photometric stereo in Sec. 7. Note that unlike related work on learning-based features [30, 37, 31], we choose not to perform spatial aggregation in the network, due to the lack of high-quality, large-scale databases of 3D objects with complex spatially-varying appearance.

6. Our Network

6.1. Input/Output

The main input to the network is a physical grayscale lumitexel at a 3D surface point corresponding to a particular pixel, representing the maximum amount of information one can acquire with an illumination multiplexing setup. The parameterization of a lumitexel is the same as the lighting layout (Fig. 1). The other input is the view specification of the current image, represented as $[\cos(\theta), \sin(\theta)]$ for continuity, in which θ is the rotation angle of the turntable along its axis. The purpose of adding this extra input is to guide the network to learn global features that are view invariant. The output is a high-dimensional feature vector. The extension to handle RGB input is described in Sec. 7.

6.2. Architecture

Our network consists of two main branches that are designed to exploit different aspects of the input photometric information. Each branch produces a separate feature vector. The two vectors are then combined with a linear fully-connected (fc) layer to generate the final feature. A graphical illustration is shown in Fig. 3.

The first branch is an intensity-sensitive one. Its first layer is a linear fc one, whose weights correspond to the lighting patterns during acquisition. This layer encodes the physical lumitexel into a small number of measurements. These measurements, along with the view specification, then go through 9 fc layers. Finally, a normalization layer produces a unit feature vector as output for training stability, as common in feature learning literature [26, 35].

The other branch is designed to be insensitive to the intensity of measurements. Overall its structure is similar to the first branch, with two major exceptions. First, there is an additional normalization layer right succeeding the measurements, to encourage the network to focus on photometric cues that are not related to input intensities. Second, the input view specification is added 5 layers after the first normalization one, to allow sufficient processing of the measurements.

Our initial attempt is to use the intensity-sensitive branch only, as it is a straight-forward, end-to-end architecture. However, the resultant features hardly exhibit variations along the azimuth angle, as visualized in Fig. 4. Adding the intensity-insensitive branch helps increase the discrimination power of features (Fig. 4) and reduce the network loss at the same time (Sec. 8.2).

Note that to physical realize the lighting patterns and to prevent degradation in the presence of training noise (Sec. 6.4), we normalize the weights in the first fc layer of each branch that correspond to a pattern, similar to [15]. During acquisition, each pattern is split into a positive and a negative one, and scaled by the inverse of the weight of the
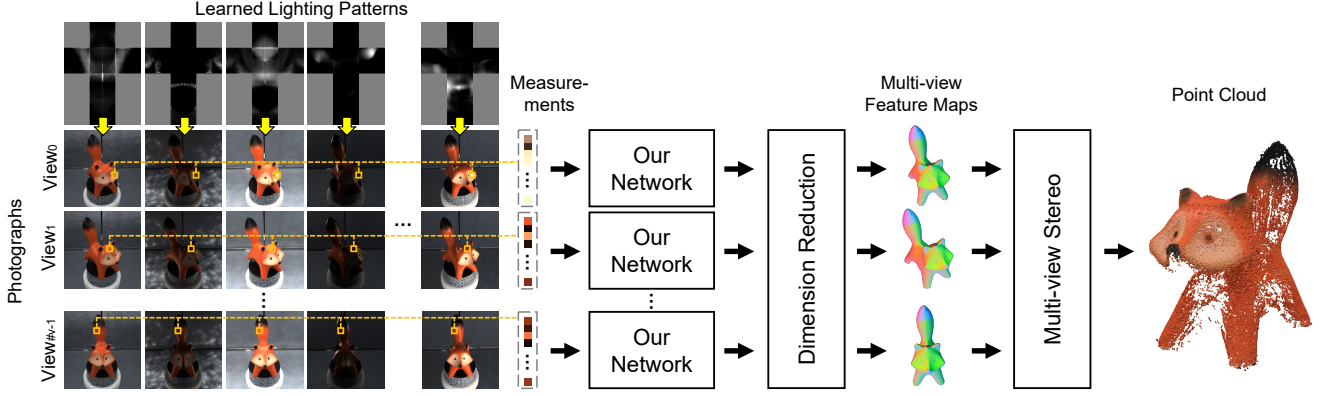
Figure 2: Our pipeline. First, we take photographs of a 3D object under learned lighting patterns at multiple views. For **each pixel** at each view, our network transforms its photometric measurements to a high-dimensional feature vector. Next, dimension reduction is performed on all transformed features across different locations and views, resulting in multi-view feature maps. Finally, we send these maps as input to a multi-view stereo technique for geometric reconstruction.
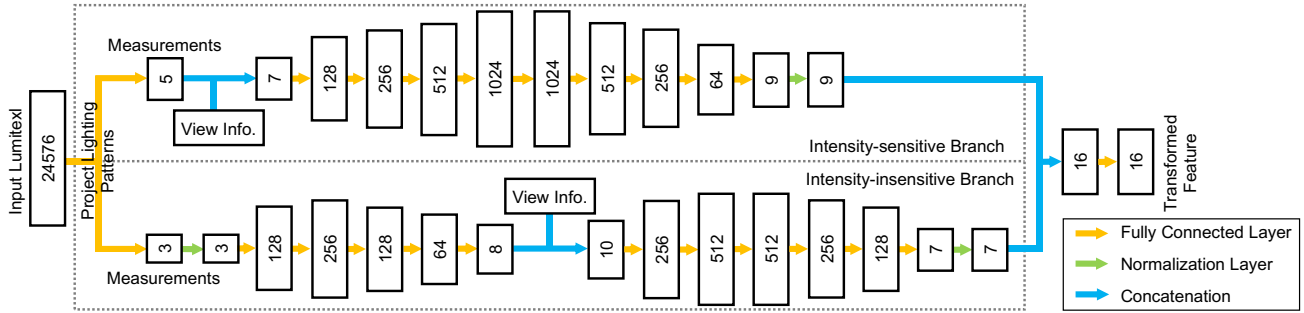

Figure 3: The architecture of our neural network, consisting of a measurement-intensity-sensitive branch and an insensitive one. For each branch, the input physical lumitexel is converted into a small number of measurements, with a linear fc layer that corresponds to the lighting patterns used in acquisition. The measurements, along with the view information of the current photograph, are transformed via a series of non-linear fc layers, before normalized to a unit feature vector. Finally, a linear fc layer combines the results from the two branches to output the final feature.

maximum absolute value, to fit in the range of $[0, 1]$.

6.3. Loss Function

Ideally, features for multi-view stereo should have the following properties: (1) the features of the same 3D point at different views are invariant; (2) the features of two different points are sufficiently discriminative; (3) and it is efficient to compare two features for matching.

Towards these goals, we define the following loss function to automatically learn to generate features embedded in a Euclidean latent space:

$$L = L_{\text{main}} + \lambda L_{\text{reg}}. \qquad (4)$$

Here L_{main} is the term to enforce the distinctiveness and invariance of features. It is modified from the E_1 term pro-
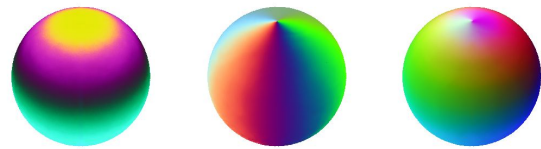

Figure 4: Visualization of learned features, computed on a synthetic ball with a homogeneous BRDF. For visualization, the original high-dimensional features are first projected to 3D via PCA and then adjusted to the range of $[0, 1]$ on a per-channel basis. From the left to right, the features from the intensity-sensitive / insensitive branch, and our network that combines the two.

posed in L2-net [31] as:

$$L_{\text{main}} = -\frac{1}{2} \left(\sum_a \log s_{aa}^{\text{col}} + \sum_a \log s_{aa}^{\text{row}} \right), \qquad (5)$$

where the related terms are defined as:

$$s_{ij}^{\mathrm{row}} = \frac{\exp\left(d_{ij}\right)}{\sum_{a=1}^{k}\exp\left(d_{ia}\right)}, \quad s_{ij}^{\mathrm{col}} = \frac{\exp\left(d_{ij}\right)}{\sum_{a=1}^{k}\exp\left(d_{aj}\right)}. \quad (6)$$

Note that d_{ij} is an entry in a feature distance matrix, constructed as follows. For each of k training points, we randomly sample related generative parameters (Sec. 6.4), and two visible view specifications. Then we compute the corresponding lumitexels at the two sampled views for each point according to Eq. 3, resulting in $\{c_1^1, c_1^2, c_2^1, c_2^2, ..., c_k^1, c_k^2\}$. Here the superscript denotes the view, while the subscript represents the point. Next, we transform these lumitexels using the current network into a set of feature vectors, as $\{h_1^1, h_1^2, h_2^1, h_2^2, ..., h_k^1, h_k^2\}$. Following [31], we define $d_{ij} = ||h_i^1 - h_j^2||_2$. This absolute Euclidean distance is converted to a relative one in Eq. 6. Note that by minimizing L_{main}, we are essentially maximizing d_{ij} for $i \neq j$ and minimizing d_{ii}, which correspond to the distinctiveness and invariance properties respectively.

We add a regularization term $L_{\mathrm{reg}} = \sum_w \|w\|_2$ to the last layer to avoid overfitting, where w is a weight in that layer. We use $\lambda = 3$ in all experiments.

6.4. Training

Similar to previous work on differential reflectance acquisition [14, 15], we synthesize the training lumitexels by randomly sampling parameters for evaluating Eq. 3 based on the GGX BRDF model, including the geometric parameters (the 3D position and the local frame), the material parameters (diffuse/specular albedos and anisotropic roughnesses), and a view specification θ. We use θ to transform the sampled position and local frame, which simulates the rotation of the turntable during acquisition. Also, we multiply each measurement with a Gaussian noise ($\mu = 1$, $\sigma = 1\%$) during training, to increase the robustness of the network.

7. Implementation Details

To handle RGB input photographs, for each pixel we apply the network to its measurements at each channel, and concatenate the per-channel results as a long vector. For efficient subsequent processing, dimensional reduction via principal component analysis (PCA) is performed on all transformed features across different views. The reduced dimensional feature maps are then fed to our modified version of COLMAP [24, 25], which operates on multi-channel floating-point input images.

To generalize to the input data in one state-of-the-art multi-view photometric stereo method [17], we replace the input of the network with a vector that concatenating all 96 measurements under different lighting conditions. This vector will be sent as input to both the intensity-sensitive and insensitive branches. As a result, the original fc layer that represents the lighting patterns (Fig. 3) is removed.

8. Experiments

Our network is implemented with PyTorch [21] on a workstation with an Intel Core i9-10940X CPU, 256GB memory and a GeForce GTX 2080 Ti graphics card. We first pre-train each branch for 100K iterations in an hour. After that, we train the complete network for 300K iterations in about 4 hours. The learning rate is set to 10^{-4}. For each view, we use a total number of 8 photometric measurements, with 5 for the intensity-insensitive branch and 3 for the sensitive one. This corresponds to $8 \times 2 = 16$ non-negative lighting patterns / input photographs per view (detailed in Sec. 6.2), which takes 15 seconds to capture with our setup. To scan a 3D object, we rotate the turntable to 24 equally spaced angles ($\theta_i = \frac{i}{24}2\pi$, $i = 0, 1, ..., 23$). We use a feature length of 16 for each color channel. The dimension is reduced to 4 with PCA before sending to COLMAP. It takes 1.3 second to transform 1 million features using the network, and 1.5 hours for our modified COLMAP to compute a dense 3D point cloud out of transformed feature maps.

We visualize our transformed per-pixel features at different views in Fig. 5. Our features respect the geometric variations (e.g., normal changes) as well as the material discontinuities (e.g., around the ears, eyes and nose of the fox), both of which are useful for discriminability. On the other hand, one can observe the reflection of the gap between LEDs from the photographs (e.g., on the fox's head and body). The gaps are inevitable as they are needed for camera installations. Our feature transform learns to reduce these adversarial, view-dependent appearance. For comparison, we also visualize the learned local patch descriptors of L2-Net [31], computed from photographs taken with a full-on lighting pattern to physically reduce the view/lighting variance of the appearance. We directly test their network pre-trained on the Brown dataset [4]. Note that our features are more view consistent compared with L2-Net, resulting in more correspondence matches as shown in the next subsection.

8.1. Comparisons

Multi-View Stereo. In Fig. 8, we compare the reconstruction results of a number of objects captured with the lightstage, using our framework against four alternative methods in multi-view stereo. CasMVSNet [11] is a state-of-the-art learning-based multi-view stereo technique. L2-Net [31] learns local features for 3D reconstruction, while differentiable reflectance capture [15] employs diffuse albedos and normals as invariant geometric features. For COLMAP [24, 25], we feed multi-view photographs taken with a full-on lighting pattern. In addition, we test
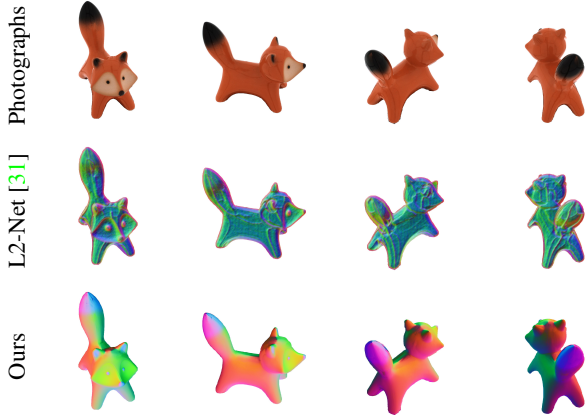
Figure 5: Feature comparisons at different views. The top row are photographs under a full-on lighting pattern. The second row visualizes the features learned by L2-Net [31] from the above photographs. Our features are visualized in the last row, which are more consistent across multiple views. All high-dimensional features are projected to 3D for visualization.
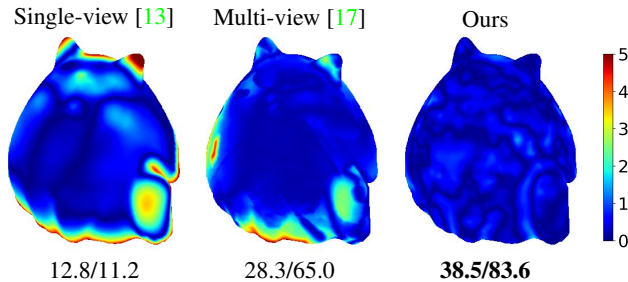

Figure 6: Comparison between single- [13] / multi-view [17] photometric stereo and our framework with a lightstage. We show color-coded geometric reconstruction errors with respect to a ground-truth model. The unit of the scale bar is mm. Quantitative errors in accuracy / completeness are reported at the bottom.

Table 1: Quantitative evaluation of the reconstruction quality of different methods (in Fig. 8), in accuracy / completeness (%) at a 1mm threshold.

	Ours	[11]	[31]	[15]	[24, 25]
Fox	**53.1/71.9**	41.1/34.2	29.4/20.0	40.2/26.9	39.9/21.3
Cat	**38.5/83.6**	33.5/23.8	29.7/08.8	28.1/13.1	38.4/05.7
Rooster	65.3/**79.8**	66.5/60.6	55.8/53.8	56.3/19.8	**70.6**/23.1
Cup	44.6/**98.2**	44.6/75.4	34.1/39.2	41.5/34.4	**51.4**/29.5

another pattern equivalent to four well-spaced point lights in Fig. 9, to introduce more shading cues for reconstruction. In all cases, existing work mainly generates 3D points
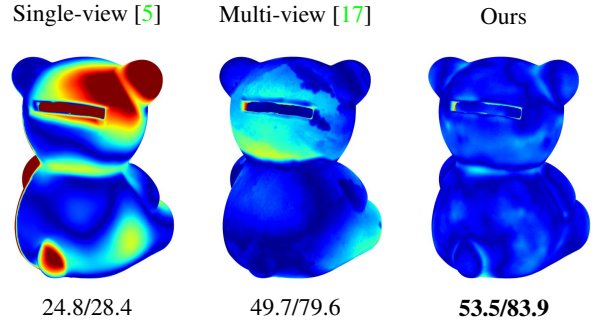

Figure 7: Comparison between photometric stereo and our framework adapted to the input from DiLiGenT-MV. We show color-coded geometric reconstruction errors. Quantitative errors in accuracy / completeness are reported at the bottom.

around regions with high spatial frequencies. For other regions, reliable correspondences cannot be established due to the lack of discriminative input information. Our method produces more complete shapes, as we fully and automatically exploit the photometric information for computing distinctive and view-invariant features (visualized in Fig. 5). Please refer to Tab. 1 for quantitative errors measured in accuracy / completeness (%) at a 1mm threshold.

Photometric Stereo. We first compare our network adapted to the data from the lightstage, against state-of-the-art single- [13] / multi-view [17] photometric stereo in Fig. 6. One hundred photographs with different illumination conditions are fed to both methods as input. To perform quantitative error analysis on the reconstructed geometry, we acquire a ground-truth shape, using a structured-lighting-based handheld industrial scanner [29], with an accuracy of 0.05mm. In comparison with related work, our reconstruction result has a lower geometric error, due to a wider coverage of reliable features correspondences (cf. Fig. 8).

Next, we compare our network generalized to the input of DiLiGenT-MV (Sec. 7) against photometric stereo on the Bear object. Our result compares favorably with [5] and [17]. For [5], the integration of normals accumulates errors in depth map generation, which are challenging to correct at a single view. For [17], their quality is sensitive to the accuracy of initial points, and its depth propagation is affected by the curvature. This experiment demonstrates our generalization ability to different input conditions / acquisition setups.

In addition, we compare in Fig. 10 the reconstruction errors of Bear between our framework and multi-view photometric stereo[17], using different number of lighting conditions sampled from DiLiGenT-MV. Our reconstruction quality is consistently higher, and does not degrade as the number of lights decreases. Fig. 11 further visualize the
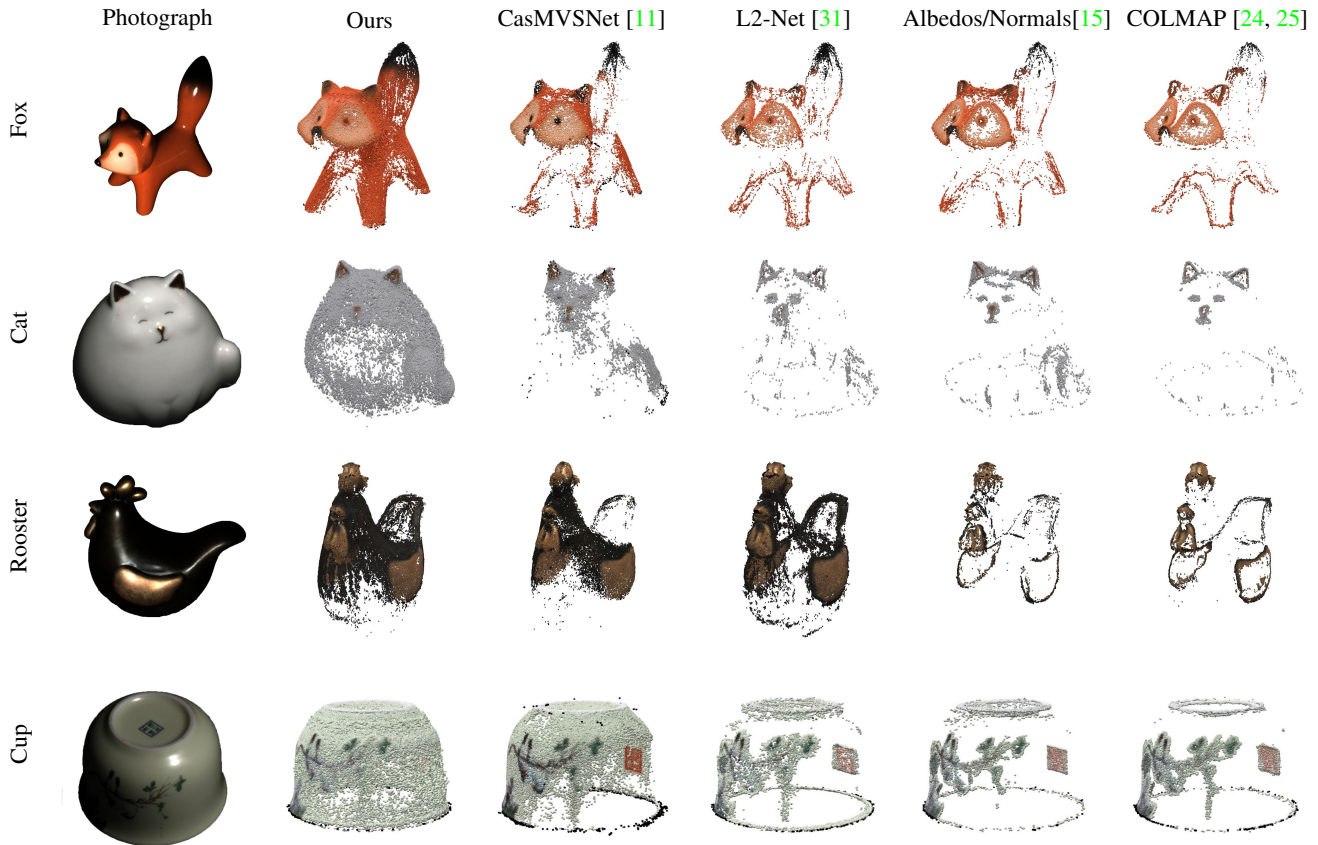
Figure 8: Comparisons with multi-view stereo methods. The first column are photographs of the physical objects. The second column to the last are geometric reconstruction results using our learned features, CasMVSNet [11], learned local features using L2-Net [31], combined results from diffuse albedo / normal maps according to [15], and COLMAP [24, 25] on images taken with a full-on lighting pattern. Corresponding quantitative errors are reported in Tab. 1.

Figure 9: Reconstruction result using COLMAP [24, 25] on photographs with 4 point lights, for more shading cues compared with a full-on pattern. The left image is a photograph, and the right geometric reconstruction. The accuracy is 34.8% and the completeness is 5.6%.
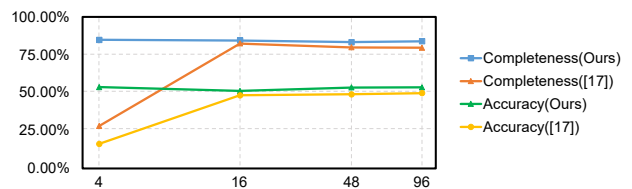

Figure 10: Comparison of reconstruction accuracy / completeness of Bear with multi-view photometric stereo [17] using different number of input lighting conditions, shown on the x axis.

results. As can be seen from the left two images, the limited lighting variations are not sufficient to predict reliable normals, even using one state-of-the-art learning-based method [5]. On the other hand, [17] struggles to identify the azimuth angle in iso-contours, resulting in shape distortions. In comparison, our reconstruction captures the major

geometric features of the original object, demonstrating the ability to efficiently exploit even the highly limited geometric information available in the input.

Moreover, we test on an object with highly challenging appearance and shape in Fig. 13. The scanner fails to produce a useful result, due to the complex light transport over the object surfaces that severely distorts the struc-

Table 2: The loss L_{main} of networks / branches with different parameters. Each corresponding bandwidth is listed between parentheses.

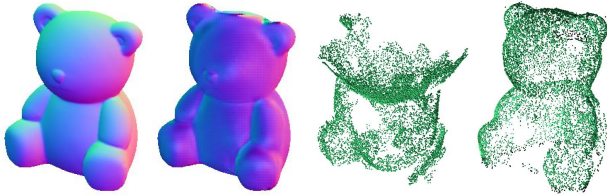| Inten.-sensitive Branch		Inten.-insensitive Branch	Our Network
56.43 (3)	53.98 (6)	55.32 (3)	11.33 (6)
56.85 (3)	53.31 (8)	53.72 (5)	10.33 (8)
53.89 (5)	53.17 (10)	53.72 (5)	9.45 (10)

Figure 11: Comparison with photometric stereo with only 4 input lighting conditions. From the left to right, the ground-truth normal map, predicted normals using [5], geometric reconstruction with [17], and our result.
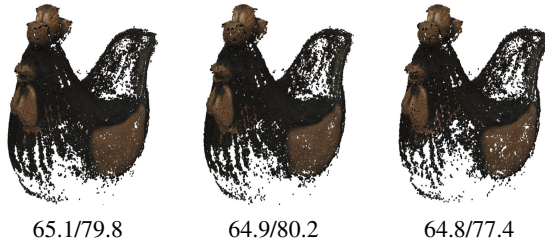

65.1/79.8   64.9/80.2   64.8/77.4

Figure 12: Impact of noise over geometric reconstruction. From the left to right, we multiply captured measurements with a a unit mean and a standard deviation of 0% / 1% / 5% Gaussian noise and show the corresponding reconstructions. Quantitative errors in accuracy / completeness are reported at the bottom.
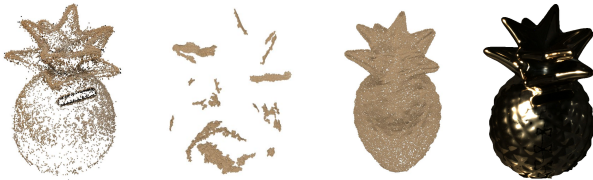

Figure 13: Geometric reconstruction on a challenging object with complex geometry and appearance. From the left to right are our result, the result from a structured-lighting-based 3D scanner [29], single-view photometric stereo [13], and a photograph.

tured pattern. For a similar reason, single-view photometric stereo [13] results in a considerably deviated depth map. As a comparison, our approach produces a 3D point cloud that captures the major geometric features of the original object.

8.2. Evaluations

We set the final feature length to 4, after balancing the reconstruction time and quality. For each branch of the network, we find that using a feature length that is roughly equal to twice its measurement number strikes a good balance between the feature distinctiveness and the computation time. We report L_{main} of networks / branches with different parameters in Tab. 2. Overall the loss decreases with the increase of the number of measurements (i.e., bandwidth). This shows that our framework can exploit the available input bandwidth to improve the feature quality. In addition, the effectiveness of adding the intensity-insensitive branch can be observed, when comparing the second column with the last, both of which have the same bandwidth at each row.

Finally, we evaluate the robustness of our features with respect to measurement noise in Fig. 12. The proposed framework produces reasonable results, even when a noise with 5 times the magnitude of its training counterpart is added, as shown in the figure.

9. Conclusion

We propose a novel framework to automatically and jointly learn to efficiently measure the physical photometric information and transform into distinctive and invariant features, which can be plugged into existing multi-view stereo pipeline for 3D reconstruction. High-quality geometric results are demonstrated over a number of daily objects with complex apperance, outperforming state-of-the-art techniques. We hope that our work will inspire interesting future work that combines multi-view and photometric cues for geometric reconstruction in a unified framework.

The proposed work is subject to a number of limitations. It does not model the appearance beyond reflectance (e.g, subsurface scattering), or explicitly handle global illumination effects like self-shadows or inter-reflections. Moreover, the current framework requires calibrated, controlled illumination. It would be promising research directions to address these limitations, to extend the idea to more general settings. We are also interested in taking other physical dimensions (e.g., polarization / hyperspectral imaging) into considerations for the extra discriminative power.

References

[1] Neil Alldrin, Todd Zickler, and David Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR*, pages 1–8, 2008. 2

[2] Neil G Alldrin, Satya P Mallick, and David J Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *CVPR*, pages 1–7, 2007. 2

[3] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *IJCV*, 72(3):239–257, 2007. 2

[4] Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *TPAMI*, 33(1):43–57, 2010. 5

[5] 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. 1, 6, 7, 8

[6] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *CVPR*, pages 1558–1567, 2017. 2

[7] M. Fiala. Artag, a fiducial marker system using digital techniques. In *CVPR*, pages 590–596 vol. 2, 2005. 2

[8] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1, 2

[9] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, pages 873–881, 2015. 2

[10] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. Shape and spatially-varying brdfs from photometric stereo. *TPAMI*, 32(6):1060–1071, 2009. 2

[11] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 5, 6, 7

[12] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *TPAMI*, 30(3):548–554, 2008. 1, 2

[13] Satoshi Ikehata and Kiyoharu Aizawa. Photometric stereo using constrained bivariate regression for general isotropic surfaces. In *CVPR*, pages 2179–2186, 2014. 6, 8

[14] Kaizhang Kang, Zimin Chen, Jiaping Wang, Kun Zhou, and Hongzhi Wu. Efficient reflectance capture using an autoencoder. *ACM Trans. Graph.*, 37(4):127:1–127:10, July 2018. 5

[15] Kaizhang Kang, Cihui Xie, Chengan He, Mingqi Yi, Minyi Gu, Zimin Chen, Kun Zhou, and Hongzhi Wu. Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Trans. Graph.*, 38(6):165:1–165:12, Nov. 2019. 2, 3, 5, 6, 7

[16] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. The digital michelangelo project: 3d scanning of large statues. In *Proc. SIGGRAPH*, pages 131–144, 2000. 1, 2

[17] 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. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020. 1, 2, 5, 6, 7, 8

[18] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1052–1061, 2019. 1, 2

[19] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *CVPR*, pages 1490–1497, 2013. 2

[20] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM transactions on graphics (TOG)*, 24(3):536–543, 2005. 1

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NIPS*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[22] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 2

[23] Joaquim Salvi, Jordi Pages, and Joan Batlle. Pattern codification strategies in structured light systems. *Pattern recognition*, 37(4):827–849, 2004. 1, 2

[24] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 5, 6, 7

[25] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 5, 6, 7

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 3

[27] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances. In *ECCV*, pages 455–468. Springer, 2012. 2

[28] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *CVPR*, pages 3707–3716, 2016. 2

[29] Shining3D. EinScan Pro 2X Plus handheld industrial scanner. https://www.einscan.com/handheld-3d-scanner/2x-plus/. 6, 8

[30] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Learning local feature descriptors using convex optimisation. *TPAMI*, 36(8):1573–1585, 2014. 2, 3

[31] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *CVPR*, pages 661–669, 2017. 2, 3, 4, 5, 6, 7

[32] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 Papers*, pages 1–11. 2009. 1, 2

[33] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, EGSR'07, page 195–206, Goslar, DEU, 2007. Eurographics Association. 3

[34] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980. 1, 2

[35] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *ICCV*, pages 2840–2848, 2017. 3

[36] Shihao Wu, Hui Huang, Tiziano Portenier, Matan Sela, Daniel Cohen-Or, Ron Kimmel, and Matthias Zwicker. Specular-to-diffuse translation for multi-view reconstruction. In *ECCV*, pages 183–200, 2018. 2

[37] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, pages 4353–4361, 2015. 2, 3

[38] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, pages 1592–1599, 2015. 2

[39] Zhenglong Zhou, Zhe Wu, and Ping Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *CVPR*, pages 1482–1489, 2013. 2