Confidence-Aware Photometric Stereo Networks Enabling End-to-End Normal and Depth Estimation for Smart Metrology

Yahui Zhang, Ru Yang, and Ping Guo, *Member, IEEE*

***Abstract*—The acquisition of geometric 3-D information is crucial for ensuring quality standards and monitoring procedures in various manufacturing applications. Photometric stereo is an established technique in computer vision to recover 3-D surfaces of objects. However, existing photometric stereo methods mainly focus on normal estimation of objects, without considering the depth estimation. On the other hand, current methods tend to prioritize accuracy while overlooking the confidence of predictions, which holds valuable information within the industry. In this article, we propose a deep learning-based photometric stereo system, consisting of hardware implementation, dataset generation, and algorithm design, to reconstruct 3-D information of physical objects, represented by normal and depth maps. In terms of the proposed algorithm, a coarse-to-fine network is introduced to improve the performance by exploiting the relationship between initial normal and depth predictions. Furthermore, the pixel-wise confidence associated with predictions is also estimated without requiring the ground truth, making a contribution to enhancing both performance and practicality. The experimental results on our synthetic dataset and real samples demonstrate the effectiveness of the proposed method on both normal/depth and confidence estimation.**

***Index Terms*—Deep learning, normal and depth estimation, photometric stereo, pixel-wise confidence.**

I. Introduction

MODERN manufacturing has been significantly enhanced by the advancement in in situ smart metrology techniques. Smart metrology plays an important role in real-time monitoring for process optimization and quality control during manufacturing processes [1], [2]. Compared to traditional dimensional metrology, 3-D measurement techniques have drawn more attention recently, as they provide intricate structural information of physical objects in 3-D space. These advanced measurement techniques will benefit the development of many fields, such as reverse engineering [3], additive manufacturing [4], and robotics [5].

Current 3-D measurement methods can be generally categorized into two types: scanning-based and image-based metrology. Scanning-based metrology involves the use of a scanning device to capture the geometry of a physical object and convert it into a digital 3-D model. The commonly used methods include laser scanning [6], coordinate measuring machine [7] and computed tomography [8]. Although the scanning-based methods provide high levels of accuracy in relevant applications, they come with some drawbacks, such as high costs and time-consuming measurements. In addition, laser scanning techniques are sensitive to reflective surfaces, leading to inaccurate measurements. The acquired data typically consist of point clouds or discrete measurements, necessitating further postprocessing, and analysis to obtain the desired outcome.

Due to the development of deep learning techniques, image-based 3-D measurement has drawn more interest. Image-based measurement utilizes the snapshots captured by single or multicameras to perform metrology, making it convenient for real-world applications. Monocular depth estimation has made significant progress [9], [10], which takes a single image as the input to predict depth information of the whole scene at the pixel level. Different from the 2.5-D representations provided by the depth estimation, neural radiance fields (NeRF)-based methods [11] offered a 3-D surface of scenes from multiview images and synthesized them in novel views. Based on image-based rendering, NeRF-based methods were optimized by minimizing the disparity between the rendered images using estimated 3-D scenes and real images. However, these methods are tailored to specific scenes and demand extensive data acquired from cameras positioned at various locations, a condition often impractical in manufacturing settings.

Another kind of image-based metrology includes deflectometry [12], structured light reconstruction [13], and photometric stereo [14]. These methods reconstruct the surface of a physical object by receiving information from the light reflection of the object, such as projected fringe patterns and reflections from different illumination conditions. Compared to structured light reconstruction and deflectometry, photometric stereo captures finer details and is more flexible to get accurate reconstruction

Received 14 September 2023; revised 30 January 2024 and 16 July 2024; accepted 10 October 2024. Recommended by Technical Editor P. Neto and Senior Editor M. Indri. This work was supported by National Science Foundation under Grant EEC-213363, Grant CNS-2229170, and Grant CNS-2328032. *(Corresponding author: Ping Guo.)*

The authors are with the Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208 USA (e-mail: yhzhang2023@gmail.com; ruyang2018@u.northwestern.edu; ping.guo@northwestern.edu).

Code is available at https://github.com/aiml-nu/Confidence-Aware-PS.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TMECH.2024.3481196.

Digital Object Identifier 10.1109/TMECH.2024.3481196

for complex surfaces. Besides, unlike NeRF, photometric stereo methods are not shape-specific and generalize well to different shapes. Therefore, in this work, we focus on photometric stereo methods to recover surfaces of objects.

Existing photometric stereo methods mainly focus on surface normal reconstruction rather than depth estimation. To get depths, traditional surface normal integration methods are usually used. However, surface normal integration is an ill-posed problem and is often limited to continuous surface features [15]. Depth information is arguably more important in manufacturing-related applications [16], [17]. It motivates this work to incorporate depth information into the reconstruction process to enhance photometric stereo with a more complete and accurate 3-D representation of the target object. Recently, Yang et al. [18] proposed a deep learning-based photometric stereo method for normal and height estimation under point-light conditions, but the interrelationship between the surface normal and height information has not been fully exploited. Different from current work [9], [10], [18], we propose to implicitly exploit the physical relationship between surface normal and depth to improve further the accuracy.

On the other hand, normal and depth estimation involve dense predictions, aiming to provide detailed predictions of normals and depths at each pixel in images. Nevertheless, ensuring precise measurements for every pixel poses a significant challenge. In such scenarios, the associated confidence in the measurements becomes crucial, as it helps identify accurate/inaccurate predictions from estimated normals and depths. Unfortunately, existing learning-based methods tend to prioritize accuracy while overlooking the confidence of the predictions, leading to a gap in their practical utility [19].

To address the abovementioned issues, we propose a confidence-aware photometric stereo network to estimate the normal and depth maps of 3-D objects as well as to provide the corresponding pixel-wise confidence levels for the prediction. The proposed networks adopt a coarse-to-fine refinement approach to exploit the relationship between surface normals and depths, aiming to further improve the prediction accuracy. Specifically, the initial normal and depth maps are predicted using UNet architectures [20] with residual connections. Then, the initial normals and depths are fed into the proposed refinement module along with the raw image inputs to leverage the initial guess estimates and their physical relationships.

Furthermore, the confidence maps are predicted without knowing the ground truth. The confidence serves two crucial purposes: i) it provides an indication of the reliability of the predictions, enabling informed decision-making and risk assessment in subsequent industry processes and ii) it further improves prediction accuracy, by implicitly prioritizing the regions of the objects involving more reflection information under lights, while de-emphasizing invalid regions where there is minimal or no change in light intensity during the network training. The predicted confidence acts as a metric to assess the reliability of the output generated by the proposed method.

To train and validate the proposed method, a synthetic dataset was customarily generated based on physics-guided information. Images were rendered from a single camera view under different lighting conditions, consistent with our proposed setup. The dataset consists of 10 objects from the Blobby shape dataset [21], as well as 15 objects with various shapes sourced from the Internet. The performance of our network is further validated and demonstrated through physical experiments.

This article presents a deep learning-based photometric stereo system including hardware and algorithm, aiming to provide a comprehensive solution that bridges the gap between theoretical research and practical applications for 3-D measurement in manufacturing area. The overall contributions can be summarized as follows.

1) An end-to-end deep learning approach is proposed to estimate both the surface normal and depth maps following the photometric stereo principles.
2) Instead of separating normal and depth estimation tasks, a coarse-to-fine network is designed to improve prediction accuracy by implicitly leveraging the physical relationships between normal and depth maps.
3) A confidence map framework is designed in the network design by a unified loss function. The confidence map can provide new insights into the prediction uncertainties (which separates this work from most in the field), but also enhance the prediction accuracy by minimizing the uncertainties in the loss function. The predicted confidence offers crucial insights and serves as an indicator of the measurement reliability.
4) The proposed method was extensively evaluated on our synthetic dataset and physical experiments. Experimental results show that our method achieves superior performance in both normal and depth estimation, while the confidence map correlates well with the prediction accuracy.

II. Related Work

In this section, we present an overview of recent advancements in the fields of photometric stereo, depth estimation, and confidence prediction. Methods for 3-D surface reconstruction are typically classified into single-view and multiview approaches. In this context, our emphasis is placed on the single-view method.

A. Learning-Based Photometric Stereo

Traditional photometric stereo, first proposed by Woodham [22], aims to recover the surface normal of objects based on the change of light intensities. The assumption is that the target has Lambertian surfaces, i.e., the observed intensities of the surface are the same under various viewing angles. However, there are few objects with Lambertian surfaces in real-world applications. To bridge the gap, deep learning techniques have been developed to handle specular reflection in photometric stereo. DPSN [14] is a pioneer photometric stereo method by using deep neural networks to deal with the non-Lambertian surface problem. To solve the disorganized and random number of inputs, Ikehata [23] introduced observation maps as the intermediate representations between each pixel of the inputs and the corresponding surface normal. Instead, Chen et al. [24] proposed to utilize the max-pooling operation to concatenate the features extracted from each image based on the same

encoder. In this approach, the sequence and number of inputs do not need to be predefined. Logothetis et al. [25] introduced a pixel-wise training manner based on the observation map to prepare samples, aiming to extend the variation but keep the same computational cost for data generation.

Although the aforementioned methods have achieved promising results on normal estimation, the depth prediction, which provides useful structural information for shape analysis, is often neglected. Different from normal estimation, depth estimation is sensitive to the textureless objects and may suffer from the noise under various lights in photometric stereo scenario. To this end, our work tries to leverage normals to enhance depth estimation via an implicit manner.

B. Depth Estimation

In real-world applications, the depth of objects holds significant importance alongside surface normals. Deep learning-based methods have achieved significant progress in monocular depth estimation [10], [26]. Similar to photometric stereo methods, monocular depth estimation involves a single camera to predict depths. These methods extract both low-level features such as edges, texture, and intensities, as well as high-level features including shapes, sizes, and positions, to infer the depth information.

Current photometric stereo methods [27], [28] attempted to predict the depth map in a data-driven manner by enforcing the consistency of predicted normals and depths. However, these approaches involve extra computational costs and may suffer from misaligned correspondence due to the limited resolution. Li et al. [29] proposed to reconstruct the new image based on decoupled surface normal, depth, and other rendering information from input images. Through the minimization of the difference between reconstructed images and input images, the predicted surface normal and depth maps were optimized. However, the above methods rely on the accuracy of estimated normal maps for the depth estimation.

In this work, we propose an end-to-end approach to predict the surface normal and depth of objects directly from raw input images. Furthermore, a coarse-to-fine refinement strategy is introduced to improve the performance by leveraging the initial estimates and implicitly exploiting the physical relationship between the surface normal and depth maps.

C. Confidence Prediction

In real-world applications, it is essential to consider not only the accuracy of the reconstructed surface but also the confidence level associated with the predictions. The confidence level provides valuable information on measurement uncertainty and guidance to further manufacturing steps. However, there are few works that take the confidence of predictions into consideration specifically in the context of photometric stereo. Recently, Kaya et al. [30] leveraged the Bayesian neural network to provide the confidence for multiview photometric stereo. However, the confidence provided in [30] is binary and only concerns the uncertainty of model parameters, not including measurement uncertainties.

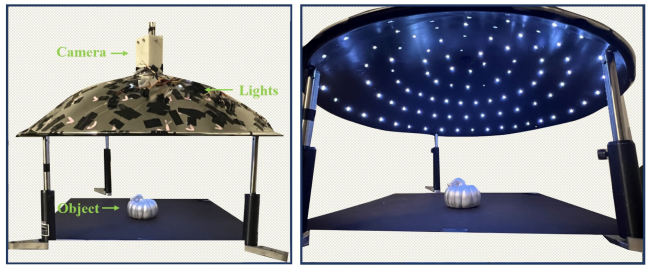


Fig. 1. Setup of photometric stereo in this study. A single camera and 96 LED lights are employed.

Instead, the use of confidence information is commonly applied in the fields of 2-D/3-D keypoint detection [31]. The keypoints are represented by the heatmaps following a Gaussian distribution, centered at the location of each keypoint. The position in the predicted heatmaps with the largest values is determined as the keypoint, where the values indicate the probability of the estimated keypoint. Inspired by keypoint detection, we propose to predict the pixel-wise confidence levels of surface normal and depth estimations in photometric stereo. Different from the task of keypoint detection, there is no ground truth available for supervising the estimation of confidence maps. In the proposed method, we introduce a novel strategy to implicitly guide the optimization of confidence estimation by combining confidence levels and estimation errors in a single loss function. The confidence prediction aims to enhance the performance of predicted normals and depths, while also addressing the challenge of lacking explicit supervision for confidence maps during training.

In contrast to existing photometric stereo methods [18], [27], [28] that neglect confidence information, we propose to predict confidence associated with normal and depth predictions and utilize predicted confidence to further improve the accuracy.

III. Method

We propose a photometric stereo system based on deep learning technique to measure the 3-D shape of objects. This system comprises of three components: photometric stereo setup implementation, physics-informed dataset generation, and deep learning algorithm design. For algorithm design, a confidence-aware photometric stereo network is proposed to achieve i) end-to-end reconstruction of accurate 3-D information of objects with the representation of surface normal and depth and ii) confidence-aware prediction with an explicit pixel-wise confidence estimation of the predictions. The overall framework is illustrated in Fig. 2. In Section III-A, we provide detailed explanations of the photometric stereo setup employed in our method. To predict the normal and depth maps of objects, our approach adopts a coarse-to-fine strategy, which is elaborated in Sections III-B and III-C, respectively. In order to enhance the performance of reconstructed surfaces and incorporate the confidence level of predictions, our method integrates confidence maps into a unified framework, which is optimized during the same training

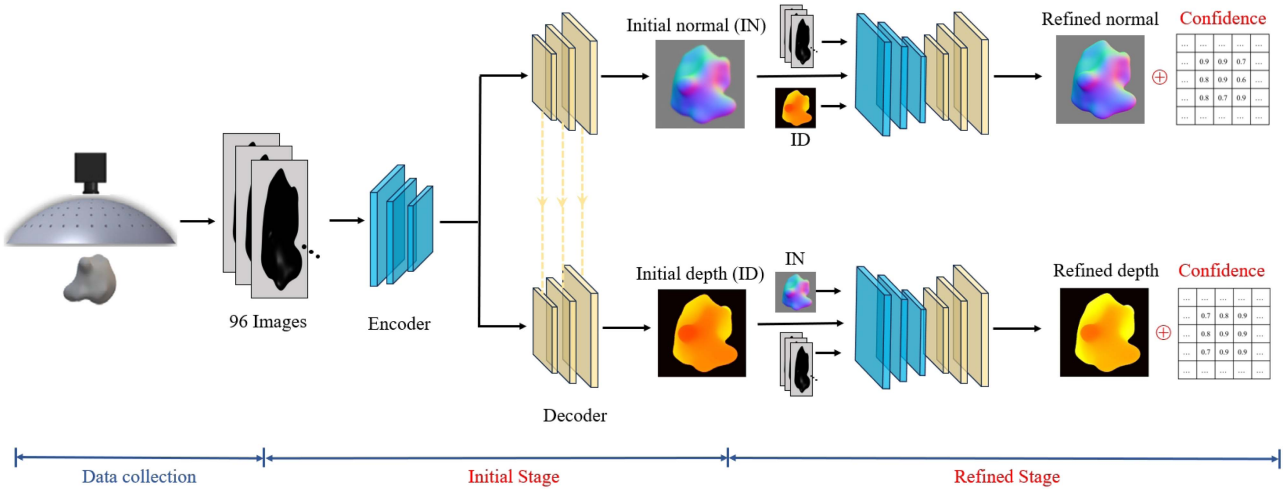


Fig. 2. Overview of the proposed method. 96 images captured under different lighting conditions are taken as the input to an encoder. Two decoders are used to predict the initial normal and depth maps, respectively. Then, the initial predictions combined with input images are fed into an autoencoder to refine normal and depth predictions. The confidence levels associated with predictions are also predicted. The feature exchange is employed in the initial stage and removed in the refined stage. The masks are used in the framework.

process with the surface normal and depth maps, as discussed in Section III-D.

A. Photometric Stereo Setup

Our photometric stereo setup is illustrated in Fig. 1. A complementary metal oxide semiconductor camera, with a focal length of 160 mm and a sensor resolution of 2048 × 1088 px, is mounted along the center of a dome shell. The diameter and focal lengths of the dome shell are 609.6 mm and 152.4 mm. A total of 96 light-emitting diodes (LEDs) are installed on the dome shell at five height levels. Particularly, an Arduino is utilized to send sequential signals to the shift registers embedded on the printed circuit board, which in turn control the lighting of each individual LED. This setup allows us to programmatically manage the illumination sequence of all 96 LEDs, ensuring they light up one at a time in a predefined order. This sequential process allows us to acquire a comprehensive set of images with diverse lighting variations, enabling a thorough analysis of the object's surface properties. Please note that we did not design a specific order to turn ON each LED light. During the comparison with current methods, all methods utilized the same data under identical LED lighting conditions for training and evaluation purposes.

To prepare the dataset, we began by calibrating both the camera and the LED positions following the calibration procedures detailed in DPPS [18]. This calibration process ensures accurate alignment and positioning. Subsequently, a synthetic dataset was generated based on the calibrated information. For further details on the dataset generation, please refer to Section IV-A. The same calibrated setup was also used to perform physical measurements as detailed in Section IV-F.

B. Initial Normal and Depth Estimation

We adopt a UNet architecture [20] followed by two decoder branches to predict normal and depth maps, respectively. Residual connections are employed in the Unets to preserve the details in the inputs and latent features. Particularly, the features obtained from the normal estimation branch, responsible for normal map estimation, are concatenated with the features from the depth estimation branch, which handles depth map estimation. This concatenated representation is used to predict the initial depth map. The motivation behind this design is from the observation that depth estimation is often more challenging in photometric stereo. The reason is that normal estimation is more directly and accurately captured through changes in pixel intensity across images. By incorporating features from normal estimation, the depth estimation branch utilizes additional information to improve the accuracy of the initial depth estimation. The normals serve as an intermediate supervision for depth estimation, allowing for better guidance and optimization during training.

In contrast to previous approaches [14], [32], we design the input representation by concatenating the captured 96 images into a single tensor with dimensions $\mathbb{R}^{96\times H\times W}$, rather than using the standard format of $\mathbb{R}^{96\times D\times H\times W}$. Here, H and W represent the height and width of the images, respectively, while D denotes the color channels (typically set to 1 for grayscale or 3 for RGB images). This offers the advantage of reducing computational costs while maintaining performance.

To train this module, we use the reverse Huber (berHu) loss [33] to minimize the difference between coarse predictions and ground truth normal/depth maps. The berHu loss is defined as follows:

$$L(\boldsymbol{x},\boldsymbol{x}^*)=\begin{cases}\sum_{i=1}^{N}|x_i-x_i^*|, & |x_i-x_i^*|\le t\\ \sum_{i=1}^{N}\frac{(x_i-x_i^*)^2+t^2}{2t}, & |x_i-x_i^*|>t\end{cases} \tag{1}$$

where N indicates the total number of samples and t is the threshold and a hyperparameter. By using the berHu loss, the model can effectively handle outliers and encourage a more robust estimation of the normal and depth maps. This loss function assigns linear loss for smaller errors to ensure smoothness, while

using quadratic loss for larger errors to penalize the outliers, leading to faster convergence. The loss function is defined as follows:

$$L_{\text{init}} = \lambda_n L(\boldsymbol{n}_{\text{init}}, \boldsymbol{n}^*) + \lambda_d L(\boldsymbol{d}_{\text{init}}, \boldsymbol{d}^*) \tag{2}$$

where $\boldsymbol{n}_{\text{init}}$ and $\boldsymbol{d}_{\text{init}}$ indicate the initial predicted normal and depth maps. λ is the weight to make a trade-OFF between different loss items.

C. Normal and Depth Refinement

We are motivated by the observation that estimated normal maps excel at capturing fine details in local regions, while predicted depth maps tend to be more accurate in global regions [9], [15]. We propose a mutual refinement approach to benefit from the strengths of each network. Specifically, we combine the initial normal and depth estimation with the 96 raw input images and feed them as the inputs to another UNet architecture. This refinement process focuses on improving the normal map estimation in global regions. Similarly, we employ a separate UNet that takes the initial depths, initial normals, and raw 96 input images as inputs to enhance the accuracy of the depth estimation in local regions. By utilizing this mutual refinement strategy, we can effectively leverage the complementary information between the two estimates, leading to improved overall performance.

In the refined module, we employ two separate UNets with residual connections to predict final normal and depth maps. Unlike the design of the coarse module, there is no information exchange between these two UNets because they learn distinct mappings to obtain the final predictions. As we know, the normal map can be approximately derived from the depth map by calculating the gradient. Hence, the refined normal estimation branch focuses on learning the mapping from the depth to the normal, i.e., derivative relationship. In the refined depth estimation branch, the mapping from the normal to the depth, i.e., integration relationship, is learned. Given that the two branches learn different mappings, we design two separate UNets without feature sharing. This design allows each UNet to capture the specific characteristics and dependencies required for the corresponding task, leading to more accurate predictions.

To train the refined module, the berHu loss between final predictions and ground truths are minimized. The loss function is

$$L_{\text{refine}} = \lambda_n L(\boldsymbol{n}, \boldsymbol{n}^*) + \lambda_d L(\boldsymbol{d}, \boldsymbol{d}^*) \tag{3}$$

where $\boldsymbol{n}$ and $\boldsymbol{d}$ denote the refined normal and depth predictions.

D. Confidence-Aware Estimation

The proposed method predicts both the surface normal and depth maps and the corresponding pixel-wise confidence map $\boldsymbol{c} \in \mathbb{R}^{H \times W}$ for the predictions, where H and W are the height and width of predicted normal and depth maps. The confidence maps contain values ranging from 0 to 1, indicating the confidence level of the predicted normal/depth maps. It provides an estimation of how much we can trust the predictions at each pixel. It is worth noting that our confidence map prediction is given without requiring knowledge of the ground truth, making it useful for practical implementation in manufacturing metrology applications.

The confidence maps are intended to reflect the reliability of the predictions, with lower confidence values assigned to erroneous regions and higher values assigned to accurate ones. There are two situations in which the confidence values tend to be lower: i) Dark regions of objects regardless of illumination conditions in the input images. In the absence of changes in light intensity, the photometric stereo method may struggle to accurately estimate the surface normal. Although deep learning-based methods have a strong ability to learn the mapping relationship between inputs and targets, predictions in these dark regions may still be less accurate unless the test samples have similar shapes to the training samples. ii) The regions where the designed model fails to handle due to its inherent limitations, such as boundary edges and extremely steep slopes. These errors arise from the uncertainties associated with the design of the model and are challenging to eliminate entirely. To avoid the overfitting problem, the weights of prediction errors in those regions should be optimally lowered to facilitate convergence.

To this end, we propose to predict the confidence map implicitly by optimizing the difference between the scaled predicted normal and depth maps and the ground truth. Let us denote the loss function $L(\boldsymbol{n}, \boldsymbol{n}^*)$ for the normal map and $L(\boldsymbol{d}, \boldsymbol{d}^*)$ for the depth map. The designed optimization function for the confidence estimation is defined as follows:

$$\begin{aligned} &\min\ L(\boldsymbol{n} \odot \boldsymbol{c_n}, \boldsymbol{n}^* \odot \boldsymbol{c}_n), L(\boldsymbol{d} \odot \boldsymbol{c}_d, \boldsymbol{d}^* \odot \boldsymbol{c}_d) \\ &\max\ \boldsymbol{c}_n, \boldsymbol{c}_d \end{aligned} \tag{4}$$

where $\odot$ indicates the Hadamard product; $\boldsymbol{c}_n$ and $\boldsymbol{c}_n$ denote the corresponding confidence map for normals and depths.

In (4), we aim to minimize the error associated with the normal $L(\boldsymbol{n}, \boldsymbol{n}^*)$ and the depth $L(\boldsymbol{d}, \boldsymbol{d}^*)$ with confidence-aware scale factors $\boldsymbol{c}_n$ and $\boldsymbol{c}_d$. To avoid the problem of trivial solutions that the network will assign all zeros to the confidence map, we introduce an additional objective function to maximize the corresponding confidence maps $\boldsymbol{c}_n$ and $\boldsymbol{c}_d$. By including the loss term of confidence maps, we encourage the model to refine the predicted confidence maps along with the normal and depth maps. This enables the model to implicitly learn the relationship between the predicted maps and their corresponding confidence values, further enhancing the accuracy and reliability of the predictions.

According to (3) and (4), the unified loss function in the refinement stage is then given by

$$\begin{aligned} L_{\text{refine}} &= \lambda_n L_n + \lambda_d L_d \\ L_n &= L(\boldsymbol{n}, \boldsymbol{n}^*) + L(\boldsymbol{n} \odot \boldsymbol{c}_n, \boldsymbol{n}^* \odot \boldsymbol{c}_n) + \lambda_{cn}(\mathbf{1} - \boldsymbol{c}_n) \\ L_d &= L(\boldsymbol{d}, \boldsymbol{d}^*) + L(\boldsymbol{d} \odot \boldsymbol{c}_d, \boldsymbol{d}^* \odot \boldsymbol{c}_d) + \lambda_{cd}(\mathbf{1} - \boldsymbol{c}_d) \end{aligned} \tag{5}$$

where $\mathbf{1}$ represents a matrix of ones with the same spatial dimension as $\boldsymbol{c}_n$ and $\boldsymbol{c}_d$.

Here, the confidence predictions are implicitly optimized by the difference between predicted and ground truth

TABLE I
PARAMETER SETTINGS IN DATASET GENERATION

Parameter	Value
Base color	0.6–0.8
Metallic	1
Specular	0.8–0.9
Roughness	0.25–0.45
Num. of objects	25
Num. of lights	96
Num. of rotation	12×12
Num. of rendered images	345,600
Resolution	512×512
Depth range	$175–374mm$
Render engine	Cycles
Projection type	Perspective

depths/normals in an end-to-end manner. These predictions are provided without necessitating the availability of ground truth data during the test.

IV. EXPERIMENTS

A. Dataset Generation

A synthetic dataset was generated by Blender using the cycles render engine [34]. Blobby shape dataset [21], consisting of 10 objects, was used to render images. As the objects in Blobby set [21] shared similar shapes, we selected 15 additional objects, from the Internet (please refer to the Supplementary Material), with more complex shapes to complement the objects from the Blobby shape dataset. These additional objects exhibit diverse and intricate geometries, providing a broader range of shapes for image rendering and training purposes. The presence of these diverse shapes ensures that the model is exposed to a wider range of geometries and can effectively handle different surface structures and variations.

A physical-based render, cycles, was applied to generate the synthetic dataset. The built-in principled bidirectional scattering distribution function was used to simulate real-world materials, influenced by the base color, metallic value, specular value, and roughness in this study. The details of parameter settings are presented in Table I. As parts in manufacturing applications are usually metallic, the metallic value is set to one and the specular value ranges from 0.8 to 0.9. The values of roughness are randomly chosen from 0.25 to 0.45, controlling surface smoothness and reflection in materials and shaders. The scale of objects is randomly set to 1.0 to 1.3, resulting in a variation range up to 200 mm in the depth range. The maximum depth value is 374 mm. The value of base color ranges from 0.6 to 0.8. To avoid overfitting, we randomly added a variation ranging from -5 to 5 mm to the calibrated positions of the camera and light positions. It should be noted that as the depth variation of objects (200 mm) is not much smaller than the distance from the object to the camera (374 mm), so we adopted the perspective projection instead of the commonly used orthographic projection for rendering. In this way, the rendered images are more realistic and suitable for depth estimation. We rotated each object with

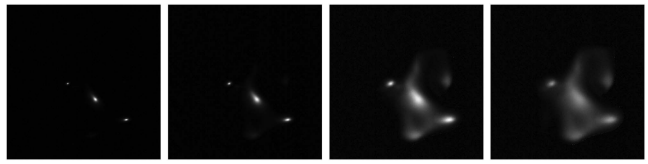


Fig. 3. Examples of rendered images of one object with various roughness (from low to high).

12×12 rotation angles. A total of 96 images were rendered for each shape input as the illumination from each calibrated light position was enabled. Therefore, the synthetic dataset consists of $25 \times 12 \times 12 = 3600$ samples. Fig. 3 illustrates exemplary rendered images from one illumination with various roughness. The training set includes 13 objects (1872 samples), while the remaining 12 objects (1728 samples) are used for evaluation.

B. Implementation Details

The designed model consists of 48.54 million parameters. The inference time is 6.7 ms for each sample. The training process involves training the coarse module first, followed by end-to-end training of the whole framework. In the first stage, L_{init} is minimized for the optimization of initial predictions and trained for 100 epochs. In the second training stage, L_{refine} is minimized, aiming to optimize the refined normal/depth predictions and confidence maps. The entire framework is then trained in an end-to-end approach for an additional 20 epochs. The learning rate is 0.00005. The batch size is 16. The hyperparameter weights are set as follows: $\lambda_n = 1, \lambda_d = 5, \lambda_{cn} = \lambda_{cd} = 0.1$. The threshold of the berHu loss t is set to 0.2. Experiments are conducted on a single NVIDIA Quadro RTX 8000 GPU.

C. Evaluation Metrics

For depth estimation, we adopt i) absolute error (Abs-err) and ii) matching error (Match-err) to calculate the difference between predicted and ground truth depth maps

$$\text{Abs-err} = \frac{1}{N}\sum_{i=1}^{N}|d_i - d_i^*| \tag{6}$$

$$\text{Match-err} = \frac{1}{N}\sum_{i=1}^{N}|d_i' - d_i^*| \tag{7}$$

$$d' = d \times \frac{\text{mean}(d^*)}{\text{mean}(d)} \tag{8}$$

where i and N represent the index and the number of samples, respectively; d, d', and d^* indicate predictions, scaled predictions, and ground truths, respectively. Abs-err directly reflects the quality of the depth predictions, while Match-err avoids the impact of the scale on errors and can better describe the geometric shape of objects.

TABLE II
EXPERIMENTAL RESULTS OF NORMAL ESTIMATION ON OUR SYNTHETIC DATASET

Method	MAE (°)	Δ	Acc05	Acc10	Acc15
DPPS [18]	13.16	61.85%	43.01%	71.87%	80.56%
NASDE [28]	8.51	41.01%	54.46%	80.81%	86.98%
Transformer [35]	6.52	23.01%	56.68%	82.07%	91.87%
Baseline	8.66	42.03%	54.43%	80.69%	86.83%
Baseline+R	6.17	18.64%	47.80%	85.95%	93.34%
Baseline+R+C	5.02	-	63.87%	87.96%	93.60%

[a] B, R, and C denote baseline, the proposed refinement module and confidence maps. Baseline+R+C is our proposed method with full components.
[b] Δ represents the percentage improvement of our method (ours) compared to each method.

TABLE III
EXPERIMENTAL RESULTS OF DEPTH ESTIMATION ON OUR SYNTHETIC DATASET

Method	Abs-err (mm)	Δ	Match-err (mm)	Δ
DPPS [18]	13.75 (6.86%)	30.69%	8.82 (4.41%)	34.24%
NASDE [28]	12.35 (6.12%)	22.83%	7.27 (3.64%)	20.22%
Transformer [35]	10.40 (5.70%)	8.37%	6.29 (3.14%)	7.80%
Baseline	12.18 (6.10%)	21.76%	7.38 (3.69%)	21.41%
Baseline+R	10.69 (5.35%)	10.85%	6.40 (3.20%)	9.38%
Baseline+R+C	9.53 (4.77%)	-	5.80 (2.90%)	-

Values in parentheses represent the ratio of error to depth range in the dataset.

For normal estimation, we use the mean angular error (MAE) as the evaluation metric

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\arccos(n_i \cdot n_i^*) \times \frac{180^\circ}{\pi} \tag{9}$$

where n and n^* represent the predicted and ground truth normal maps.

D. Comparison

To evaluate the performance, we compare our method with three state-of-the-art methods: DPPS [18], NASDE [28] (that directly enforces the consistency between normals and depths), and transformer-based method [35]. The baseline represents the proposed method without i) the refinement module (referred to as R) and ii) confidence information (referred to as C). Besides, we normalize the input images by the maximum grayscale values of all input images, so we keep the consistency of the light intensities for all inputs.

1) Overall Performance: The experimental results of normal and depth estimation are presented in Tables II and III, respectively. For the normal estimation, the MAE and the percentage of predictions with errors less than 5°, 10°, and 15° are used in the assessment. These metrics provide insights into the accuracy and reliability of the predicted surface normals across the test set.

The following can be seen.

1) Our baseline outperforms DPPS with an improvement of 4.50° for the normal estimation and 1.57 mm for the depth estimation. It demonstrates that the performance of the photometric stereo method benefits from the berHu loss and normalization preprocessing of the inputs.
2) The results indicate the propose method significantly outperforms current methods DPPS and NASDE with an improvement of 61.85% (DPPS) and 41.01% (NASDE) for normal estimation and 30.69% (DPPS) and 22.83% (NASDE) for depth estimation, and slightly outperforms transformer [35] with an improvement of 23.01% for normal estimation and 8.37% for depth estimation. In addition, the proposed method (6.07 ms) is obviously fast than NASDE (38.52 ms) and transformer (49.17 ms), as described in the Supplementary Material.
3) As the input images are captured by one single camera, there exists inherent ambiguity in the depth dimension.

Match-err can better represent the accuracy of reconstructed surfaces by alleviating the depth ambiguity. Therefore, the Match-err is lower than Abs-err for all three methods. The proposed method outperforms current methods with the Match-err of 5.80 mm, which is around 2.90% to the depth range and 1.55% to the maximum depth values in the proposed dataset.

The qualitative comparison is illustrated in Fig. 4. The proposed method generates more accurate normal and depth maps compared to DPPS.

2) Ablation Study: In the conducted ablation study, the effectiveness of each component in the proposed method is analyzed by comparing different configurations: Baseline with the proposed refinement module (Baseline + R), and baseline with proposed refinement module and confidence maps (Baseline + R + C), i.e., the proposed method. The results are reported in Tables II and III, providing valuable insights into the impact of each component on the performance of normal and depth estimation.

1) *Comparison between three methods*: The results demonstrate that both the refinement module and the introduced confidence map contribute to the overall performance improvement.
2) *Analysis of baseline and Baseline + R*: From Table II, the Baseline + R configuration achieves superior performance in metrics, such as MAE, Acc10, and Acc15. However, it exhibits lower accuracy in the Acc05 metric. This discrepancy can be attributed to the refinement process, which primarily concentrates on enhancing the global regions of the initial normals by leveraging the advantages of the initial depths. Consequently, there may be a relative degradation in performance within the local regions of the normal map.
3) *Impact of the confidence module*: The proposed confidence module addresses the abovementioned issue. The Baseline + R + C significantly improves the accuracy from 47.80% to 63.87% in the metric of Acc05. This improvement is attributed to the incorporation of confidence information, enabling the optimization process to effectively acknowledge and account uncertainties in various regions of reconstructed objects.
4) *Differential improvement in normal and depth estimation*: The experimental results indicate that the improvement in normal estimation with the inclusion of confidence information is more significant compared to depth estimation in photometric stereo. This discrepancy is attributed to

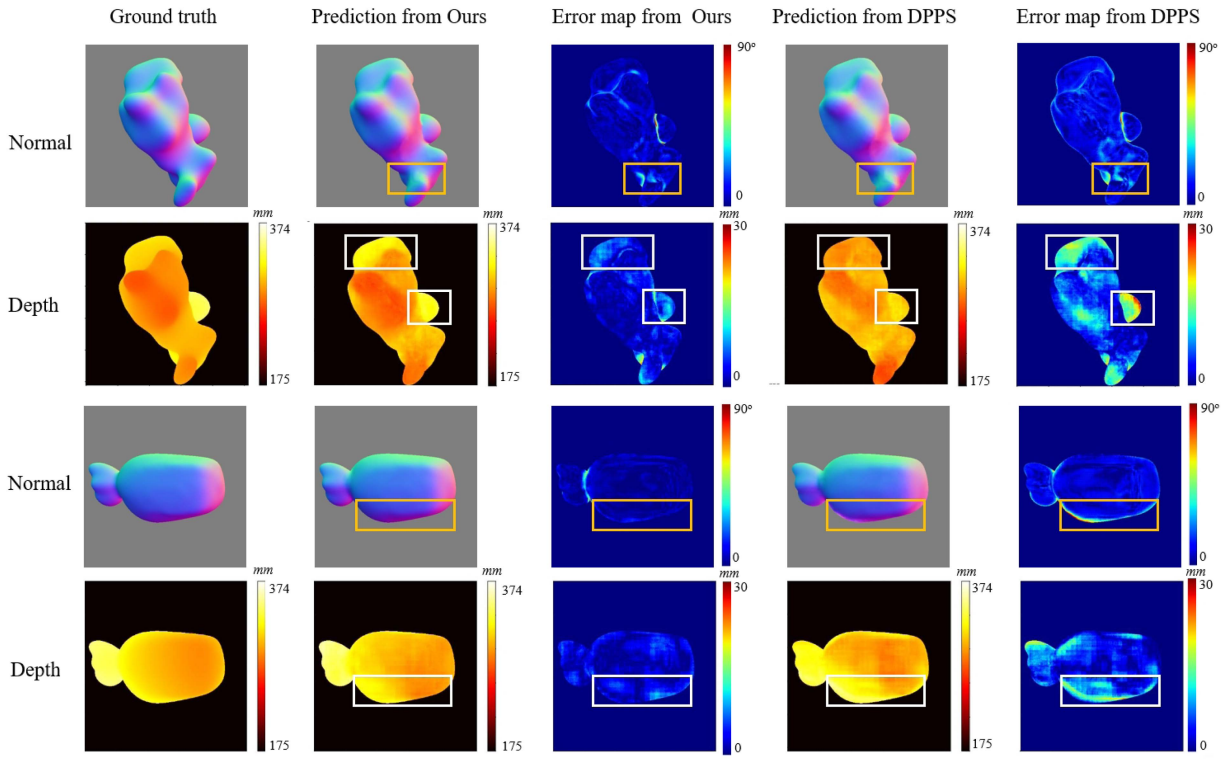


Fig. 4. Comparison of the proposed method with the existing method (DPPS [18]) for normal and depth estimation on the test set. Backgrounds are removed by the masks, rendering the values in the background meaningless.

the inherent challenges of depth estimation on integrating information across the scene under various illumination and textureless surfaces. These challenges inherently introduce higher uncertainty in depth predictions.

The qualitative comparison of the ablation study is involved in the supplementary material. By conducting this ablation study and analyzing the comparative results, the effectiveness of each component in the proposed method is verified, and valuable insights are gained regarding their impacts on the performance of normal and depth estimation.

E. Discussion

1) Validation of Confidence Map: In Figs. 5 and 6, the estimated normal/depth maps and their corresponding confidence maps are showcased. It is observed that the predicted confidence maps exhibit a close correspondence with the error maps, indicating that the predicted confidence maps effectively capture the uncertainty and provide valuable guidance about the confidence level of the predictions. Notably, the predicted confidence maps are optimized by the proposed framework without relying on predefined or ground truth confidence.

The results obtained from the ablation study, combined with the visualizations presented in Figs. 5 and 6, emphasize the significance of the predicted confidence maps in practical applications, particularly in industrial contexts. These confidence maps offer valuable insights into the reliability and accuracy of the predictions even without the requirement of ground truth. They pinpoint areas where predictions may be less reliable, enabling researchers to focus efforts on improving accuracy in those specific regions.

2) Analysis of Confidence Estimation for Depths and Normals: Compared to the predicted confidence map in the depth estimation, the estimated confidence in the normal estimation shows a better alignment with the error map, as illustrated in Figs. 5 and 6. Particularly, the confidence level is presented in the range from 0 to 1 for the normal estimation, and from 0.9 to 1 for the depth estimation.

This is because the photometric stereo method used for normal estimation is more robust compared to depth estimation from a single camera. For normal estimation, the key factor is the material properties and the change of light intensities, which can be inferred from images captured by a single camera. However, depth estimation is more challenging due to its vulnerability in textureless surfaces and susceptibility to noise under various lights, making it difficult to accurately predict depth maps. Consequently, the predicted confidence in the depth estimation may not align as closely with the error map as observed in the normal estimation.

3) Confidence Map Versus Attention Map: In this study, the predicted confidence map is different from the widely used attention map, but they share a similar concept to provide extra information to steer the prediction process. The attention map [36] focuses on assigning higher weights to relevant and representative features while reducing the influence of unrelated features. The aim is to enhance the performance of a specific task by selectively attending to important information. Instead, the confidence map is designed to estimate the confidence level of

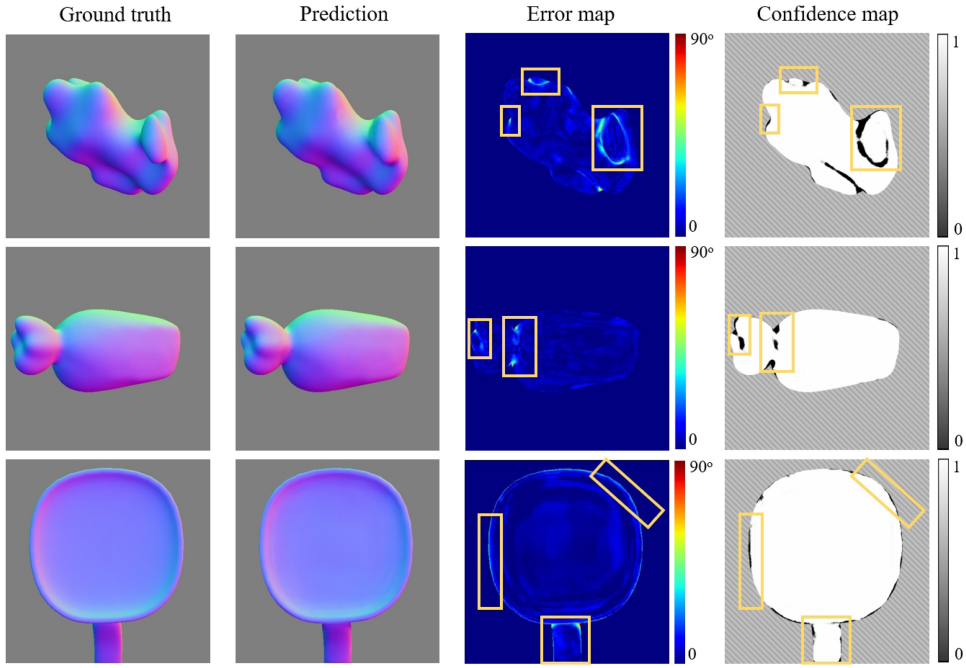


Fig. 5. Visualization of the normal map and the corresponding confidence map generated by the proposed method from the test set. The regions with smaller confidence values match well with the regions with larger errors of predictions.

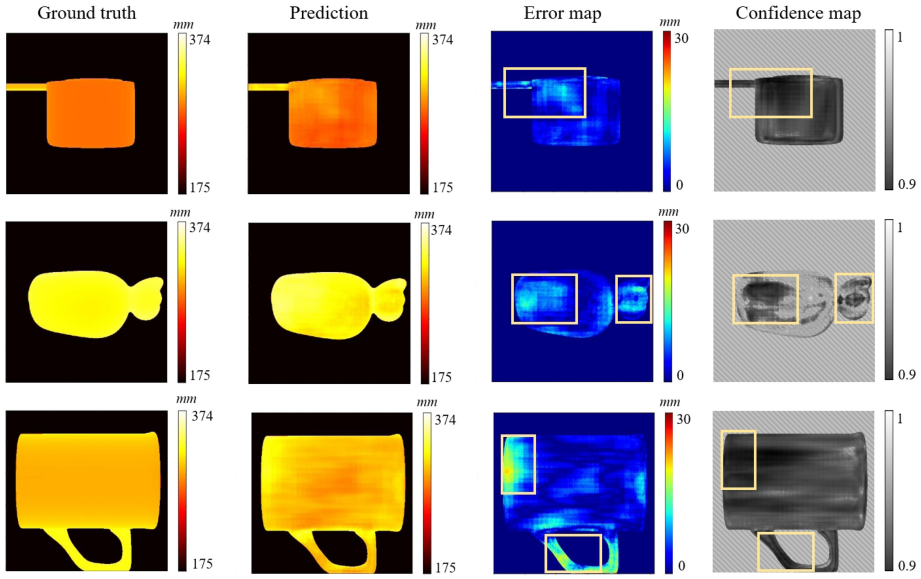


Fig. 6. Visualization of the depth map and the corresponding confidence map generated by the proposed method from the test set. The regions with smaller confidence values match well with the regions with larger errors of predictions.

predictions, which is particularly valuable by providing uncertainty information about the measurement. The confidence map is implicitly optimized by the difference between predicted and ground truth normal and depth maps. It captures the reliability of the predictions and provides valuable insights into the quality of the results. By incorporating the confidence map into the optimization process, the model can assign lower weights to predictions in regions that are invalid or where the model faces challenges due to inherent limitations, leading to an overall more accurate estimation.

F. Evaluation on Real Samples

To validate the generalization ability of the proposed method, we conducted physical experimental verifications on real objects. To capture images from real objects, we employed the

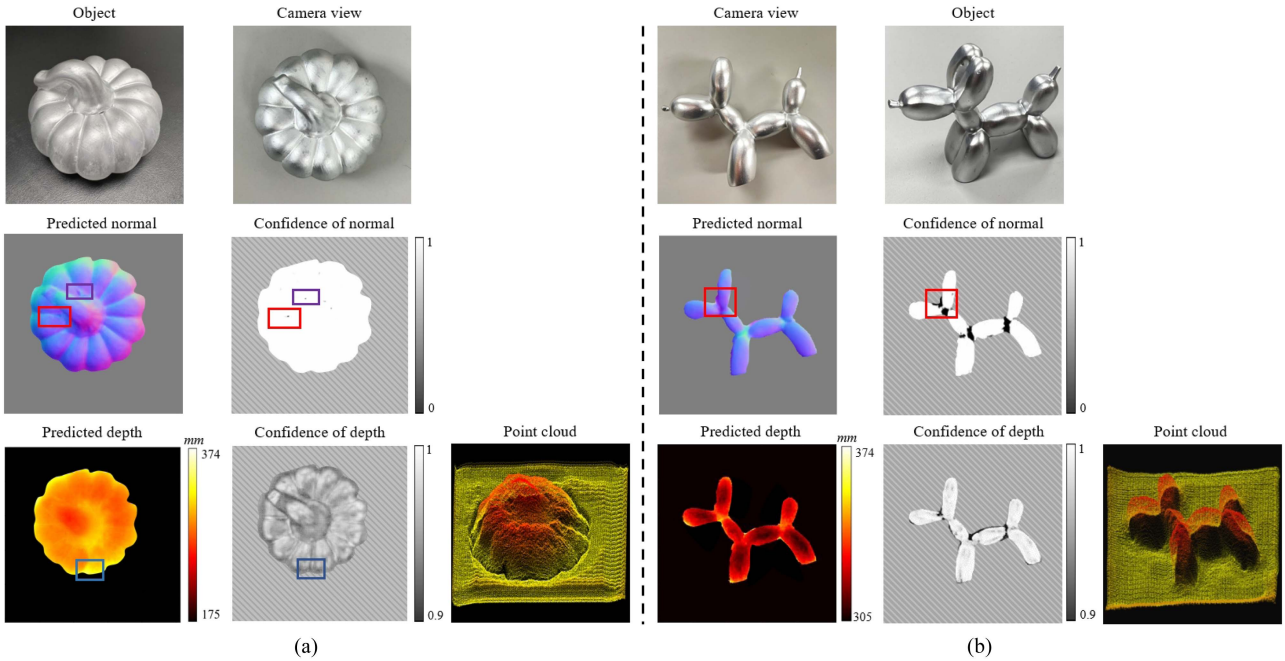


Fig. 7. Visualization of predicted surface normals, depths, and corresponding confidence maps from physical experiments of measuring (a) pumpkin sculpture and (b) bear sculpture. The rectangles with the same color indicate a consistent relationship. The predicted depth is represented as a point cloud for better visualization.

setup described in Section III-A. Specifically, two objects are chosen for this evaluation: a pumpkin sculpture and a bear sculpture, both with reflective surfaces. The visualization of predictions is presented in Fig. 7. It can be seen that visualization of the predictions demonstrates plausible normals and depths for both pumpkin and bear sculptures. Moreover, we use the results in Fig. 7(a) as an example to illustrate the relationship between predictions and corresponding confidences: i) the confidence levels in the rectangles are significantly lower than other regions, which denotes that the predicted normals in the rectangle regions are in low confidence. In fact, the visualization of the predicted normal shows two artifacts in the rectangle regions. ii) The predicted depths in the blue rectangle regions are not reasonable from the visualization due to the inconsistency with neighboring values. Correspondingly, the confidence values in the blue rectangle regions are lower than in other regions. Physical experimental verifications are presented in the supplementary video.

V. Conclusion

In this article, we present a confidence-aware photometric stereo system based on deep learning for end-to-end normal and depth estimation from images under different lighting conditions. Our method exploits the relationship between the initial estimation of normal and depth maps to further improve the accuracy of predictions by the introduced coarse-to-fine refinement framework. To get the confidence level of predictions, a novel strategy is proposed to implicitly optimize and predict the confidence maps. This enables the network to provide meaningful confidence estimates for the normal and depth maps, even in the absence of explicit ground truth for the confidence maps. Extensive experiments demonstrate superior performance achieved by the proposed method compared to the existing methods and our baselines.

References

[1] Y. Liu, W. Zhao, H. Liu, Y. Wang, and X. Yue, "Coverage path planning for robotic quality inspection with control on measurement uncertainty," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 5, pp. 3482–3493, Oct. 2022.

[2] S. E. Sadaoui, C. Mehdi-Souzani, and C. Lartigue, "Multisensor data processing in dimensional metrology for collaborative measurement of a laser plane sensor combined to a touch probe," *Measurement*, vol. 188, 2022, Art. no. 110395.

[3] Z. Geng and B. Bidanda, "Tolerance estimation and metrology for reverse engineering based remanufacturing systems," *Int. J. Prod. Res.*, vol. 60, no. 9, pp. 2802–2815, 2022.

[4] K. Zhu, J. Y. H. Fuh, and X. Lin, "Metal-based additive manufacturing condition monitoring: A review on machine learning based approaches," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 5, pp. 2495–2510, Oct. 2022.

[5] T.-H. Wang and P.-C. Lin, "A reduced-order-model-based motion selection strategy in a leg-wheel transformable robot," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 5, pp. 3315–3321, Oct. 2022.

[6] E. P. Baltsavias, "A comparison between photogrammetry and laser scanning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 54, no. 2/3, pp. 83–94, 1999.

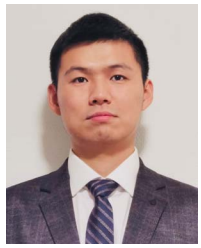
[7] M. M. P. A. Vermeulen, P. Rosielle, and P. Schellekens, "Design of a high-precision 3d-coordinate measuring machine," *Cirp Ann.*, vol. 47, no. 1, pp. 447–450, 1998.

[8] Q. Ma et al., "Autonomous surgical robot with camera-based markerless navigation for oral and maxillofacial surgery," *IEEE/ASME Trans. Mechatron.*, vol. 25, no. 2, pp. 1084–1094, Apr. 2020.

[9] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4009–4018.

[10] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "Bifuse: Monocular 360 depth estimation via bi-projection fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 462–471.

[11] F. Zhu et al., "Deep review and analysis of recent nerfs," *APSIPA Trans. Signal Inf. Process.*, vol. 12, no. 1, 2023, Art. no. e6.
[12] L. Huang, J. Xue, B. Gao, C. McPherson, J. Beverage, and M. Idir, "Modal phase measuring deflectometry," *Opt. Exp.*, vol. 24, no. 21, pp. 24649–24664, 2016.
[13] J. Zhang et al., "A convenient 3D reconstruction model based on parallel-axis structured light system," *Opt. Lasers Eng.*, vol. 138, 2021, Art. no. 106366.
[14] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Self-calibrating deep photometric stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8739–8747.
[15] Y. Quéau, J.-D. Durou, and J.-F. Aujol, "Normal integration: A survey," *J. Math. Imag. Vis.*, vol. 60, pp. 576–593, 2018.
[16] J. Lai, B. Lu, and H. K. Chu, "Variable-stiffness control of a dual-segment soft robot using depth vision," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 2, pp. 1034–1045, Apr. 2022.
[17] K. Song, J. Wang, Y. Bao, L. Huang, and Y. Yan, "A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 3, pp. 1558–1569, Jun. 2023.
[18] R. Yang, Y. Wang, S. Liao, and P. Guo, "DPPS: A deep-learning based point-light photometric stereo method for 3D reconstruction of metallic surfaces," *Measurement*, vol. 210, 2023, Art. no. 112543.
[19] G. A. da Silva, A. T. Beck, and O. Sigmund, "Topology optimization of compliant mechanisms considering stress constraints, manufacturing uncertainty and geometric nonlinearity," *Comput. Methods Appl. Mech. Eng.*, vol. 365, 2020, Art. no. 112972.
[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[21] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2553–2560.
[22] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Opt. Eng.*, vol. 19, no. 1, pp. 139–144, 1980.
[23] S. Ikehata, "CNN-PS: CNN-based photometric stereo for general non-convex surfaces," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.
[24] G. Chen, K. Han, and K.-Y. K. Wong, "PS-FCN: A flexible learning framework for photometric stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.
[25] F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla, "PX-NET: Simple and efficient pixel-wise training of photometric stereo networks," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 12757–12766.
[26] Y. Zhang, M. Gong, J. Li, M. Zhang, F. Jiang, and H. Zhao, "Self-supervised monocular depth estimation with multiscale perception," *IEEE Trans. Image Process.*, vol. 31, pp. 3251–3266, 2022.
[27] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 283–291.
[28] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2189–2199.
[29] J. Li and H. Li, "Neural reflectance for shape recovery with shadow handling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16221–16230.
[30] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool, "Uncertainty-aware deep multi-view photometric stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12601–12611.
[31] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, and J. He, "Deep learning on monocular object pose detection and tracking: A comprehensive overview," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–40, 2022.
[32] Y. Ju, Y. Peng, M. Jian, F. Gao, and J. Dong, "Learning conditional photometric stereo with high-resolution features," *Comput. Vis. Media*, vol. 8, pp. 105–118, 2022.
[33] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis.*, IEEE, 2016, pp. 239–248.
[34] "Cycles.," [Online]. Available: https://www.cycles-renderer.org/
[35] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 16269–16279.
[36] D. Wang, Y. Li, L. Jia, Y. Song, and T. Wen, "Attention-based bilinear feature fusion method for bearing fault diagnosis," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 3, pp. 1695–1705, Jun. 2023.



Yahui Zhang received the M.Sc. degree in mechanical manufacturing and automation from the Huazhong University of Science and Technology, Wuhan, China, in 2018, and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 2022.

His research focuses on human pose estimation, photometric stereo, and computer vision.

Ru Yang received the B.S. degree in energy and power engineering from Xi'an Jiaotong University, Xi'an, China, in 2017, and the Ph.D. degree in mechanical engineering from Northwestern University, Evanston, IL, USA, in 2023.

She is currently an Algorithm Engineer with KLA, Milpitas, CA, USA.

Ping Guo (Member, IEEE) received the B.S. degree in automotive engineering from Tsinghua University, Beijing, China, in 2009 and the Ph.D. degree in mechanical engineering from Northwestern University, Evanston, IL, USA, in 2014.

He is currently an Associate Professor of Mechanical Engineering with Northwestern University. From 2014 to 2018, he was an Assistant Professor with the Chinese University of Hong Kong, Hong Kong, China. His research interests include advanced manufacturing, precision engineering, and 3-D printing.