IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 March 2023, accepted 17 April 2023, date of publication 21 April 2023, date of current version 26 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3269286

RESEARCH ARTICLE

Applying Albedo Estimation and Implicit Neural Representations to Well-Posed Shape From Shading

WANXIN BAO[1], **REN KOMATSU**[1], **(Member, IEEE), HAJIME ASAMA**[1], **(Fellow, IEEE), AND ATSUSHI YAMASHITA**[2], **(Senior Member, IEEE)**
[1]Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo (UTokyo), Tokyo 113-8656, Japan
[2]Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo (UTokyo), Chiba 277-0882, Japan
Corresponding author: Wanxin Bao (baowanxin@robot.t.u-tokyo.ac.jp)

ABSTRACT We present a method that improves the accuracy of depth maps by combining albedo estimation and implicit neural representations to the well-posed shape from shading. Because the estimation of depth information from a single image is an under-constrained problem, we apply certain physical constrains to convert the ill-posed shape from shading problem to a well-posed problem. Subsequently, we construct an image irradiance equation wherein the surface parameter representing albedo is estimated using a learning-based encoder-decoder network. By solving the equation using implicit neural representations, we can obtain a depth map of the original image. The proposed method achieves an accuracy of depth estimation from a single image with the mean absolute error (MAE) of 0.1510 and root mean square error (RMSE) of 0.1768, indicating superior performance to that of existing methods. Both simulation and real experiments have been carried out to verify the effectiveness of the proposed method.

INDEX TERMS Depth estimation, intrinsic image decomposition, implicit neural representations, shape from shading.

I. INTRODUCTION

Scene depth estimation plays a significant role in computer vision by enhancing the perception and understanding of the surrounding environment. Consequently, a wide range of applications, such as autonomous driving, virtual reality, and robotic navigation require depth estimation [1]. Active depth estimation techniques use lasers or structured light to obtain point clouds and estimate depth maps [2]. Although these depth maps are highly accurate, the required measuring equipment is usually large, making these techniques inconvenient to employ in narrow environments. To investigate environments with collapsed bricks at disaster sites or to inspect the internal conditions of a machine, the environment is usually dark and narrow. Usually, there is no light source in these environments. As a result, a camera with a self-contained light source is suitable for dark and narrow environments.

Given a single image, shape from shading (SfS) can recover the depth map by analyzing the intrinsic components of the image, such as albedo, surface normal, and light source. This analysis is not simple owing to the concave or convex ambiguity [3]. By employing physical constraints and certain assumptions, the ill-posed SfS problem can be converted to a well-posed problem [4], [5]. However, the surface parameter albedo cannot be obtained in advance and have to be assigned arbitrarily, thereby degrading the accuracy of estimated results.

In recent years, many learning-based SfS methods have been developed and demonstrated high accuracy. Supervised learning with a Convolutional Neural Network (CNN)-based encoder-decoder has been used to train a network that jointly predicts reflectance, depth, and light conditions [6]. Liu et al. explore the independence between shading and reflectance and propose an unsupervised intrinsic image decomposition

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

framework to estimate intrinsic components [7]. Learning-based methods can also achieve good accuracy and jointly estimate multiple intrinsic components using a network. However, these methods usually require a large amount of ground truth training data to train the network.

In this study, depth estimation by implicit neural representations and albedo estimation for well-posed SfS is proposed. We leverage the physical constrains to formulate the SfS problem and utilize implicit representations to parameterize the partial differential equation. Analyzing existing SfS methods, we find two primary disadvantages that contribute to suboptimal accuracy. Specifically, the surface parameter albedo is not estimated in advance and methods for solving the image irradiance equation do not perform sufficiently. The objective of this study is therefore to improve the accuracy of depth estimation using SfS.

Our methodology is summarized as follows. First, the ill-posed SfS problem is converted to a well-posed problem based on certain assumptions and constraints. This guarantees the only solution to this problem. Because albedo is a vital yet initially unknown parameter, it is estimated using a learning-based method. Subsequently, the estimated albedo map is combined with the well-posed SfS. Finally, the image irradiance equation is solved using implicit neural representations. This work is an extension of our previous work presented at a conference, wherein the image irradiance equation is solved using implicit neural representations, however, the albedo parameter remains unknown [8].

The main contributions of this study are as follows.

- We estimate the albedo map of the original image, and apply it to a well-posed shape from shading problem.
- We solve the well-posed shape from shading problem using implicit neural representations.

This paper will proceed as follows: In Section I, we present a background of our research and summarize the primary methods and contributions of this article. In Section II, we review prior studies as they relate to our own. In Section III, we explain our proposed method in detail. We present a series of simulation and real experiments in Section IV. We compare our proposed method with existing approaches in Section V. Finally, Section VI concludes our paper and presents limitations and potential future directions of research.

II. RELATED WORK

In this study, we convert the ill-posed SfS problem to a well-posed problem and build a partial differential equation, wherein learning-based intrinsic image decomposition is introduced to estimate the albedo parameter. We employ implicit neural representations and a gradient descent-based method to solve the partial differential equation. In the following subsections, we review three lines of work in SfS, intrinsic image decomposition, and gradient descent-based methods to solve partial differential equations.

A. SHAPE FROM SHADING

SfS reconstructs the shape of an original image from a gradual variation of shading therein [9]. It is to recover the depth map by analyzing the intrinsic components of a single image, including albedo, surface normal, and light source from a single image.

SfS is introduced by Horn who first derive an equation describing the relationship between the shape of a surface and its corresponding brightness [10]. Most SfS methods follow two steps to solve the problem: first, proposing assumptions and constructing an image irradiance equation; second, solving the equation using numerical methods [4].

Early SfS approaches are based on the following assumptions:

- The object surface is a Lambertian surface
- The light source is located at infinity
- The camera model is orthographic

A Lambertian surface is a surface that emits the same brightness irrespective of the observer's perspective, thereby representing an ideal diffuse surface. Based on the above assumptions, the image irradiance equation can be built and numerically solved. Although these assumptions simplify the complex shading and imaging process, the reconstructed results have poor accuracy because it is impossible for an object surface to satisfy perfect diffuse reflection. Furthermore, most cameras employ the pinhole design, and the light source is generally near the object, rather than at infinity.

To improve the accuracy of solutions to the SfS problem, Lee and Kuo propose a more realistic imaging model, that uses a perspective camera projection, as well as a light source located at the optical center [5]. Using an endoscope camera, Okatani and Deguchi propose a notion of an equal distance contour and obtain the equation for this contour [11]. Ikeda applies a linear approximation to a hybrid reflectance map composed of Lambertian model and Phong model [12], [13]. Ahmed and Farag use a linear combination of the Lambertian model and Ward model, which considers diffuse reflection and specular reflection of the object surface, with the objective of attaining higher imaging precision [14]. Fan et al. use the Cook-Torrance BRDF reflectance model to express the hybrid surface, and introduce variational formulation to solve the image irradiance equation [15]. Cao et al. applies SfS in medical imaging by combining it with stereo vision to reconstruct heart models [16].

In summary, the SfS problem exploits physical constraints to express the relationship between the shape of a surface and its corresponding brightness. Subsequent to building an equation, various methods for solving the equation also affect the accuracy of the estimated shape.

B. INTRINSIC IMAGE DECOMPOSITION

Intrinsic image decomposition (IID) is traditionally described as the problem of decomposing an image into two layers: albedo, invariant color of the material, and shading, produced by the interaction between light and geometry [17]. The main relationship between IID and SfS is the estimation of intrinsic properties by analyzing a single image. In recent years, deep learning techniques have been broadly applied to improve the accuracy of this process. According to learning strategy, IID methods may be categorized as weakly, fully, or self-supervised.

Weakly supervised methods require human observations relating to perceptions of materials and illumination in an image. Two datasets, IIW and SAW, have been generated based on human observations, where regions corresponding to similar albedo or assigned shading labels are respectively contained [18], [19]. These judgments are expressed via sparse sets of pairwise comparisons and accompanied by confidence scores representing the weighted human disagreement rate (WHDR) [20]. Kovacs et al. estimate the source of shading gradient using a CNN with a classifier and subsequently apply the local prediction within a classical Retinex formulation to estimate the intrinsic components [19].However, the human observations required by weakly supervised methods may cause inaccuracies in the dataset. In addition, human annotations alone are insufficient for training a direct regression approach, likely because such annotations are sparse and derived from a few thousand real images [21].

The fully supervised strategy entails learning from labeled data. The use of a labeled dataset to train a model is a common strategy in machine learning. For each labeled image, the loss function is penalized if the estimated intrinsic component does not conform to the corresponding ground truth label. Narihira et al. utilize the ground truth of the MPI Sintel dataset, obtaining results for both synthetic and real images, sourced from Sintel and the MIT intrinsic image dataset respectively [22]. Their network is a CNN-based architecture that takes an RGB image and directly predicts its albedo and shading. Luo et al. apply surface normal estimation to IID [23], whereas Ma et al. propose a cascaded network with two sub-networks designed for reflectance estimation and shading optimization [24]. Intrinsic image decomposition has also jointly estimated with semantic segmentation [25].

In self-supervised works, image formation loss is calculated to guarantee that the target parameters of the network effectively reconstruct the original image. Janner et al. propose a CNN-based encoder-decoder network, where a shared encoder and three separate decoders for the estimation of albedo, shape, and light conditions [6]. The shape and light conditions are utilized to train a differential shading function in another network architecture where the self-supervised strategy is applied to reconstruct the original image. Yu and Smith incorporate multiview stereo poses and depth maps for cross projection [26]. Using a differential renderer and an InverseRenderNet, which accepts an RGB image and outputs the albedo and normal, this network learns via self-supervision.

C. GRADIENT DESCENT-BASED PARTIAL DIFFERENTIAL EQUATION SOLUTION

Deep learning has revolutionized fields such as images, text, and speech recognition, which require statistical approaches to model non-linear functions of high-dimensional inputs. Using multi-layer neural networks, deep learning has proven effective in practice for numerous tasks in the aforementioned fields. One such task involves solving partial differential equations (PDEs).

In general, a neural network requires an appropriately designed loss function to solve an equation. By updating a set of network parameters, the loss function can be minimized and the solution to the equation can be obtained. The Deep Galerkin method (DGM) calculates a high-dimensional PDE using deep neural network [27]. By minimizing the squared error, this method can solve the high-dimensional Hamilton-Jacobi-Bellman PDE and Burger's equation. The Deep Ritz method (DRM) utilizes two residual connections in its network, thereby avoiding the vanishing gradient problem [28]. SIREN applies a sinusoidal function to the activation function for a multi-layer perceptron, and solves the PDE using implicit neural representations [29].

III. METHOD

In this study, depth estimation using a learning-based method for a well-posed PDE is proposed. First, we convert the ill-posed SfS problem to a well-posed problem by applying certain assumptions and construct an image irradiance equation. The unknown parameter ρ representing albedo in the equation is estimated by a learning-based method. Finally, we solve the image irradiance equation using implicit neural representations.

The main plan of this study can be divided into three phases:

- Applying assumptions to convert the ill-posed SfS problem to a well-posed problem and constructing an image irradiance PDE.
- Training a neural network to estimate albedo, and applying the learning results to the well-posed SfS problem.
- Utilizing implicit neural representations to solve the PDE.

A. CONSTRUCTION OF IMAGE IRRADIANCE EQUATION VIA WELL-POSED SHAPE FROM SHADING BASED ON CERTAIN ASSUMPTIONS

1) PERSPECTIVE CAMERA MODEL AND LIGHT SOURCE AT OPTICAL CENTER

We introduce a perspective camera model for the projection model [30]. Figure 1 illustrates the camera model with perspective projection and a point light source at the optical center is illustrated. The object's surface is represented by $S(\boldsymbol{x})$, where f is the focal length of the camera. The intersection of the light direction and image plane can be expressed as $(\boldsymbol{x}, -f)$, where $\boldsymbol{x}$ is a pixel on the image plane. The unit light source vector $\boldsymbol{L}(\boldsymbol{x})$ can be represented as

$$\boldsymbol{L}(\boldsymbol{x}) = \frac{1}{\sqrt{f^2 + ||\boldsymbol{x}||^2}} \begin{pmatrix} -\boldsymbol{x} \\ f \end{pmatrix}. \tag{1}$$

Consider the surface S representing the object or scene of interest in a given image domain Ω, parameterized with the

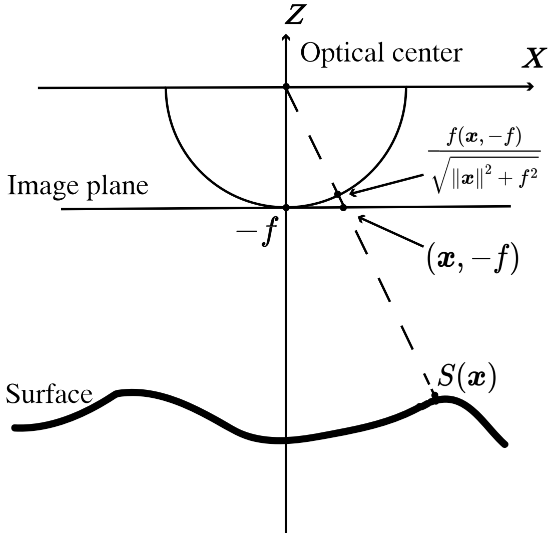


FIGURE 1. Perspective camera model. Point light source is located at optical center.

function $S\colon \Omega \to \mathbb{R}^3$ by

$$S(\boldsymbol{x}) = \frac{fu(\boldsymbol{x})}{\sqrt{f^2+||\boldsymbol{x}||^2}}\begin{pmatrix}\boldsymbol{x}\\ -f\end{pmatrix}, \tag{2}$$

where

$$\boldsymbol{x} = \begin{pmatrix}x_1\\ x_2\end{pmatrix} \in \Omega. \tag{3}$$

Ω denotes the image domain and $u(\boldsymbol{x})$ represents the depth in the projection direction. The normal vector of surface point $S(\boldsymbol{x})$ can be obtained by calculating the cross product of tangent vectors in the x_1 and x_2 directions. As a result, the normal vector $\boldsymbol{n}(\boldsymbol{x})$ for surface point $S(\boldsymbol{x})$ is described as

$$\boldsymbol{n}(\boldsymbol{x}) = \begin{pmatrix} f\nabla u(\boldsymbol{x}) - \frac{f\nabla u(\boldsymbol{x})}{f^2+||\boldsymbol{x}||^2}\boldsymbol{x}\\ \boldsymbol{x}\cdot\nabla u(\boldsymbol{x}) + \frac{f\nabla u(\boldsymbol{x})}{f^2+||\boldsymbol{x}||^2}f\end{pmatrix}. \tag{4}$$

The term $\cos\theta_i$ is the dot product between $\boldsymbol{L}(\boldsymbol{x})$ and $\boldsymbol{n}(\boldsymbol{x})$ using the change of variables $v(\boldsymbol{x}) = \ln u(\boldsymbol{x})$,

$$\begin{aligned}\theta_i &= \arccos\frac{\boldsymbol{n}(\boldsymbol{x})}{||\boldsymbol{n}(\boldsymbol{x})||}\cdot\boldsymbol{L}(\boldsymbol{x})\\ &= \arccos\frac{Q(\boldsymbol{x})}{\sqrt{f^2||\nabla v(\boldsymbol{x})||^2+(\boldsymbol{x}\cdot\nabla v(\boldsymbol{x}))^2+Q^2(\boldsymbol{x})}},\end{aligned} \tag{5}$$

where $Q(\boldsymbol{x}) = \frac{f}{\sqrt{f^2+||\boldsymbol{x}||^2}}$.

2) OREN-NAYAR REFLECTION MODEL

One of the simplest reflection models typically used in computer vision and computer graphics is the Lambertian model, which models a surface that scatters incident illumination equally in all directions. Because this is an ideal diffuse reflection model, it cannot precisely represent the surface of a real object.

This section introduces a more realistic reflection model, namely the Oren-Nayar reflection model [31]. This model is more comprehensive than the Lambertian model because it fully considers geometric and radiometric phenomena including inter-reflection, masking, and shadowing between points on the surface. The surface is assumed to comprise a collection of small V-cavities based on the microsurface theory.

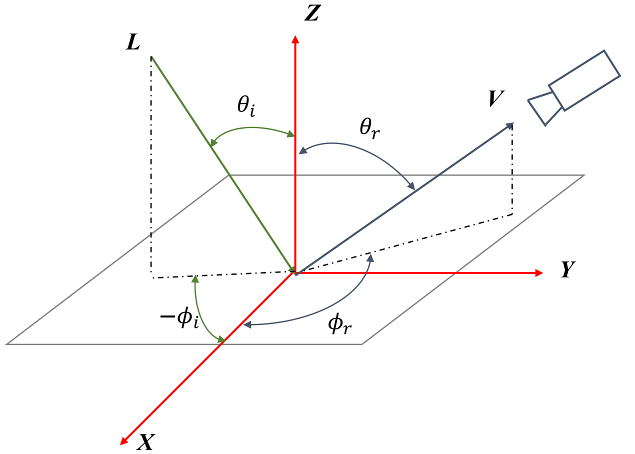


FIGURE 2. Reflection model.

Figure 2 depicts the incident and reflected angles. (θ_i, ϕ_i) is the incident direction, and (θ_r, ϕ_r) is the reflected direction.

A simplified expression for reflected radiance can be summarized as follows [31]:

$$\begin{aligned}&L_r(\theta_i, \phi_i; \theta_r, \phi_r)\\ &= \frac{\rho}{\pi}I_0\cos\theta_i \times (A + B\max[0, \cos(\phi_r - \phi_i)]\sin\alpha\tan\beta),\end{aligned} \tag{6}$$

where $A = 1-0.5\frac{\sigma^2}{\sigma^2+0.33}$, $B = 0.45\frac{\sigma^2}{\sigma^2+0.09}$, $\alpha = \max[\theta_i, \theta_r]$ and $\beta = \min[\theta_i, \theta_r]$.

The albedo ρ represents the fraction of incident energy reflected by the surface, and I_0 denotes the intensity of the point light source. The parameter σ measures the surface roughness.

Due to perspective camera model,

$$\theta_i = \theta_r = \alpha = \beta, \tag{7}$$

$$\phi_i = \phi_r. \tag{8}$$

An attenuation term $1/r^2$ is considered to further guarantee a well-posed SfS problem. r indicates depth along the projection, which is to be estimated.

As a result, the reflected radiance equation becomes

$$L_r = \frac{\rho}{\pi}\frac{I_0}{r^2}(A\cos\theta_i + B\sin^2\theta_i). \tag{9}$$

The relationship between image brightness and surface radiance is

$$E_i = L_r\frac{\pi}{4}\left(\frac{d}{f}\right)^2\cos^4\chi, \tag{10}$$

where E_i is the image irradiance, considered to be equal to the image brightness. d is the diameter of the lens, and f is the focal length. χ is the angle between the optical axis and line of sight to a surface point of a corresponding image point. Although the term $\cos^4\chi$ implies nonuniform brightness even for uniform illumination, the actual optical system is designed to correct it. As a result, we may consider image brightness to be proportional to surface radiance:

$$E_i = \lambda L_r. \tag{11}$$

If we denote $I = \frac{E_i}{\lambda I_0}$, the brightness equation becomes

$$I = \frac{1}{r^2}\frac{\rho}{\pi}(A\cos\theta_i + B\sin^2\theta_i). \tag{12}$$

3) BUILD IMAGE IRRADIACE EQUATION

Based on the perspective camera model assumption and Oren-Nayar reflection model, the image irradiance equation can be constructed as follows:

$$-e^{-2v(\boldsymbol{x})} + \frac{f^2 I(\boldsymbol{x})}{D(\boldsymbol{x}, \nabla v)} = 0, \forall \boldsymbol{x} \in \Omega, \tag{13}$$

where

$$D(\boldsymbol{x}, \nabla v) = \frac{\rho}{\pi}\left[A\frac{Q(\boldsymbol{x})}{\sqrt{F(\boldsymbol{x}, \nabla v) + Q^2(\boldsymbol{x})}} + B\frac{F(\boldsymbol{x}, \nabla v)}{F(\boldsymbol{x}, \nabla v) + Q^2(\boldsymbol{x})}\right], \tag{14}$$

$$F(\boldsymbol{x}, \nabla v) = f^2||\nabla v(\boldsymbol{x})||^2 + (\boldsymbol{x}\cdot\nabla v(\boldsymbol{x}))^2, \tag{15}$$

$$r = fu(\boldsymbol{x}) = fe^{v(\boldsymbol{x})}. \tag{16}$$

The conversion of an ill-posed SfS problem to a well-posed problem and the subsequent construction of an image irradiance equation are similar to the approach proposed by [4]. Our concept of feeding the estimated albedo map to the image irradiance equation and using implicit neural representations to solve the equation is introduced in the following subsection.

B. ALBEDO ESTIMATION USING LEARNING-BASED METHOD

To train a neural network for albedo estimation, we require a dataset that meets the following problem setting: the only light source in the environment is the point light located at the optical center of the camera. Because no such dataset exists, we generated our own dataset.

1) DATASET GENERATION

We modified the work of [6] to place a point light source at the optical center of the camera. The 3D models in this dataset were obtained from ShapeNet [33] and the dataset of intrinsic decomposition images was generated by Blender [34]. ShapeNet, which includes a subset called ShapeNetCore, encompasses 55 common object categories with approximately 51,300 unique 3D models. 3D object models provided by ShapeNet vary across a wide range of shapes and materials. The objective is to use a perspective camera to capture photos of these 3D models and generate their corresponding intrinsic images in Blender. Intrinsic decomposition images, which include depth, albedo, shape, light source, and shading attributes, are obtained from original images, where the light source information is presented with the help of a sphere. Although the point light source's location is constant at the optical center of the camera, its intensity is randomly generated.

2) SIMULATION RESULTS

Each category in the dataset includes two thousand sets for training, eight hundred sets for testing and two hundred sets for validation. We divided the datatset by 11-fold and used k-fold cross-validation in the training phase.

Figure 3 depicts the sample results of data generation.

3) ARCHITECTURE OF NETWORK TO ESTIMATE ALBEDO

This section describes the network architecture for learning albedo. The model in question has a convolutional encoder-decoder architecture, wherein mirror-link connections are employed to connect encoder and decoder of equal size [6].

The encoder comprises five convolutional layers with 16, 32, 64, 128, and 256 filters of size 3 × 3 and strides of 2. Batch normalization and ReLU activation are performed after each convolutional layer. The albedo decoder has the same size as the encoder, but uses a three-channel output. Thus, the network takes a three-channel original image as the input, and outputs the corresponding albedo map.

C. SOLVING THE IMAGE IRRADIANCE EQUATION

1) APPLYING ESTIMATED ALBEDO TO IMAGE IRRADIANCE EQUATION

The obtained albedo map corresponds to the parameter ρ in Eq. (14). By substituting the albedo map into Eq. (14), the PDE expressed in Eq. (13) can be solved with a higher accuracy than if the albedo were arbitrarily assigned. Comparison results are presented in Section IV.

2) IMPLICIT NEURAL REPRESENTATIONS TO SOLVE PARTIAL DIFFERENTIAL EQUATION

We introduce sinusoidal representation networks (SIREN) to solve the complex PDE [29]. SIREN is a simple neural network architecture that utilizes a sine function as a periodic activation function for implicit neural representations. For an equation that satisfies the form below

$$F(\boldsymbol{x}, \Phi(\boldsymbol{x}), \nabla\Phi(\boldsymbol{x}), \ldots) = 0, \Phi : \boldsymbol{x} \mapsto \Phi(\boldsymbol{x}) \tag{17}$$

a neural network that parameterizes $\Phi(\boldsymbol{x})$ to map $\boldsymbol{x}$ while satisfying the constraint in Eq. (17) can be obtained. This implicit problem formulation takes the spatial coordinate $\boldsymbol{x} \in R^m$ as the input. $\Phi(\boldsymbol{x})$ is implicitly defined by function F.

We parameterize Φ_θ as a fully-connected neural network with parameters θ, and solve the optimization problem using gradient descent. The neural network is a four-layer network with one layer of input, three hidden layers, and one layer of output. All layers are multi-layer perceptrons (MLPs) with a sinusoidal function as the activation function.

For the pixel coordinates $\boldsymbol{x}_i$ in Eq. (3) considering the image irradiance equation in Eq. (13), the loss function is

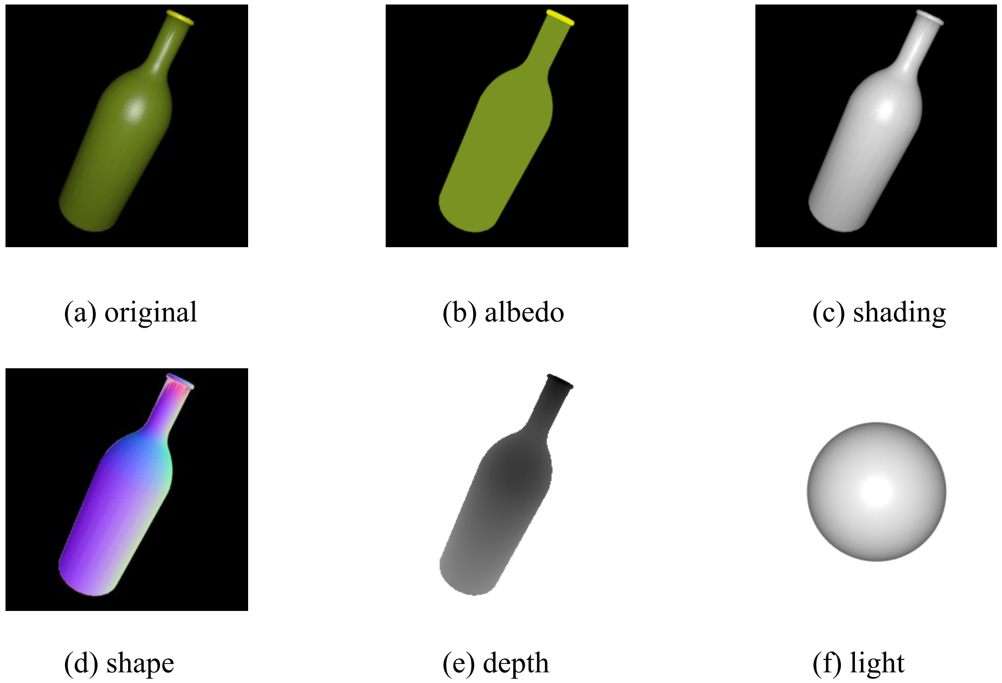


FIGURE 3. Example of dataset. One original image of a bottle and its corresponding albedo, shading, shape, depth and lighting condition maps. The lighting condition map is presented with the help of a sphere.

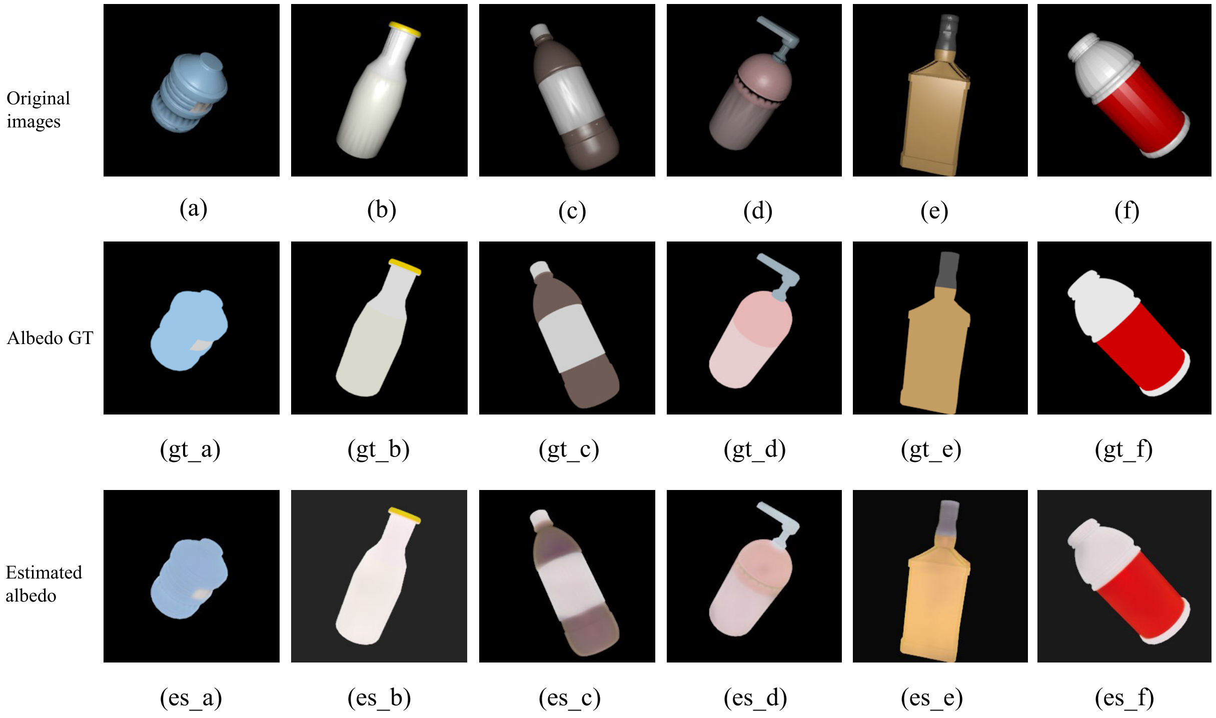


FIGURE 4. Albedo estimation results for six simulation images.

defined as follows:

$$J(\Phi) = || - e^{-2\Phi_\theta(\boldsymbol{x}_i)} + \frac{f^2 I(\boldsymbol{x}_i)}{D(\boldsymbol{x}_i, \nabla\Phi)} ||_\Omega^2. \qquad (18)$$

$J(\Phi)$ measures how well the function $\Phi_\theta(\boldsymbol{x})$ satisfies the PDE differential operator. If $J(\Phi) = 0$, then $\Phi_\theta(\boldsymbol{x})$ is a solution to the PDE.

The objective is to determine a set of parameters θ so that $\Phi_\theta(\boldsymbol{x})$ minimizes the loss function $J(\Phi)$. If the loss $J(\Phi)$ is small, then $\Phi_\theta(\boldsymbol{x})$ closely satisfies the PDE. In the proposed method, the parameters θ are updated using the ADAM algorithm, which has proven to be a very effective optimizer in machine learning [32]. ADAM yields fast and robust convergence, and its hyper-parameters have intuitive interpretations and typically require little tuning.

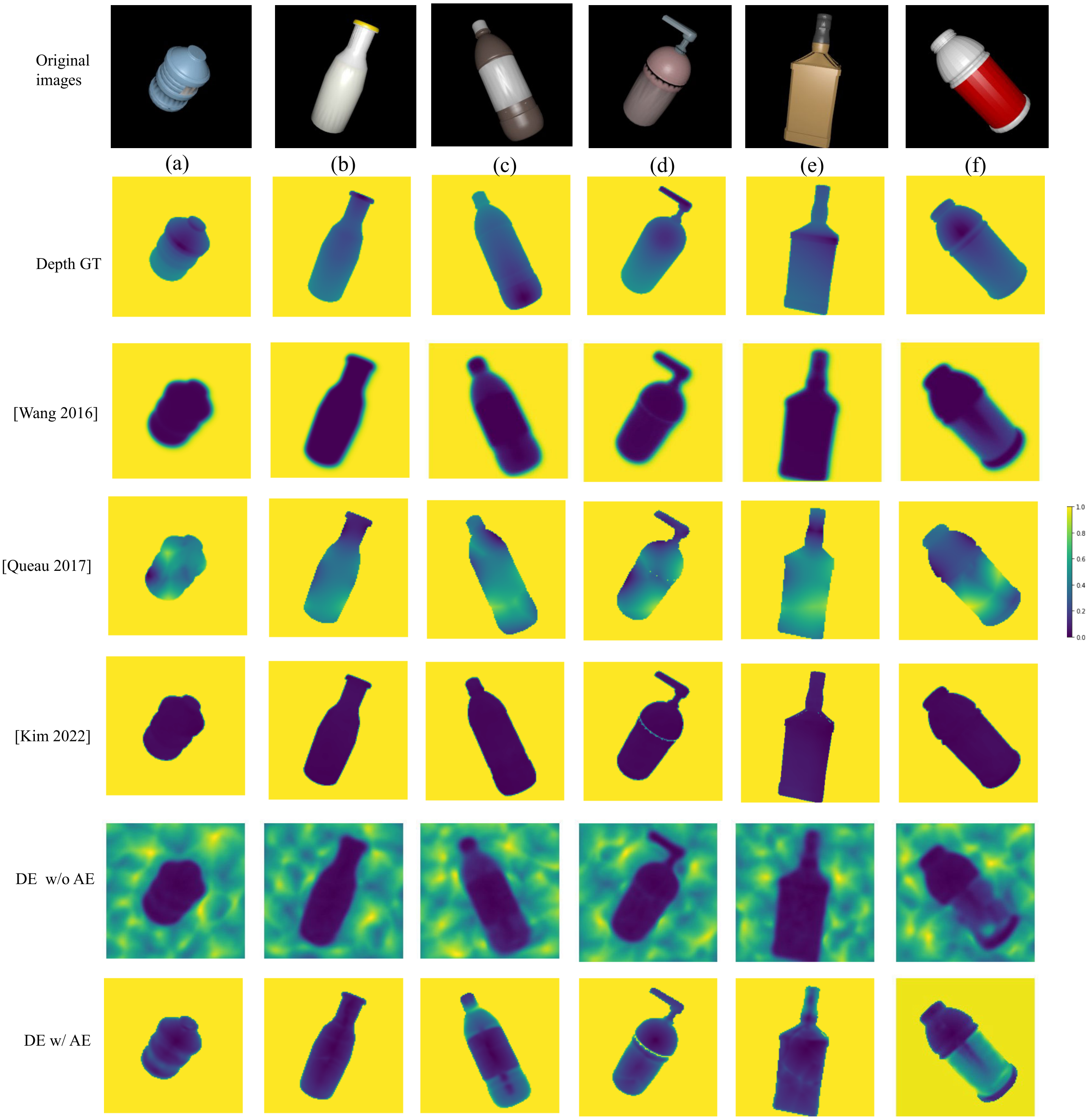


FIGURE 5. **Depth estimation of the simulation experiment.**

IV. RESULTS

To evaluate our proposed method's performance, we conducted both simulation and real experiments.

In the simulation experiment, the original images were obtained from Blender. All images were generated in an environment where the point light source is situated at the optical center of the perspective camera.

In the real experiment, images were captured in a dark environment using both an endoscope camera and a smartphone. The depth maps estimated by the proposed method and several previous approaches are presented explicitly.

A. SIMULATION EXPERIMENT

1) SIMULATION SETTING

In the simulation experiment, we conducted quantitative analysis to evaluate the performance of depth estimation using simulated images. First, a neural network was trained to learn albedo. In this step, the original images and albedo maps for fully-supervised learning were obtained from Blender [34]. Subsequently, the original images and their albedo maps were sent to SIREN to estimate the depth maps.

The network architecture was trained with dataset encompassing the bottle category. Data were allocated among training, testing and validation subsets according to a

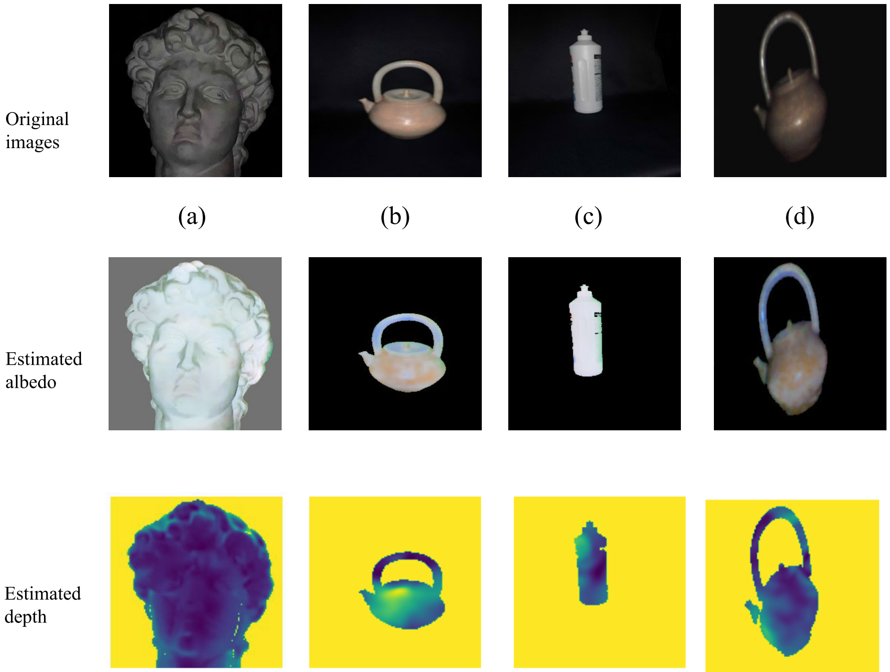


FIGURE 6. Albedo estimation and depth estimation in real experiment. (a), (b), and (c) were captured using a smartphone. (d) was captured with an endoscope camera.

2000:800:200 ratio, and k-fold cross-validation was applied in the training phase. Two thousand original images with their albedo ground truth images were used for training over fifty epochs, and two hundred sets of original images with their albedo ground truth images were utilized for validation. Six different original images selected to represent the estimated albedo maps are shown in Figure 4.

After an albedo map is obtained, it substitutes the parameter ρ in Eq. (14) and is sent to the SIREN network, which directly outputs the depth map.

Albedo estimation results for the six selected bottles are shown in Figure 4. To ensure evaluation accuracy, we repeated the experiment over five rounds. The mean absolute error (MAE) and root mean square error (RMSE) were then calculated as quantitative metrics [35]. The MAE is given by

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |\hat{D}_i - D_i|, \tag{19}$$

where $\hat{D}_i$ is the estimated value and D_i is the actual value.

The RMSE is given by

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{D}_t - D_t)^2}{T}}, \tag{20}$$

where $\hat{D}_t$ is the estimated value and D_t is the actual value.

A quantitative analysis of the albedo estimation is shown in Table 1.

According to the estimated albedo maps and quantitative analysis, the proposed method achieves highly accurate albedo estimation.

TABLE 1. MAE and RMSE of estimated and ground truth albedo for six images.

Images	MAE	RMSE
Fig.4(a)	0.0466	0.0493
Fig.4(b)	0.0778	0.0803
Fig.4(c)	0.0453	0.0653
Fig.4(d)	0.0461	0.0553
Fig.4(e)	0.1134	0.1211
Fig.4(f)	0.0715	0.0782

The depth estimation results obtained by the proposed and previous methods are shown in Figure 5, which presents the original images, depth ground truth, and estimated depth maps. A corresponding quantitative comparison is presented in Table 2, where DE represents depth estimation, and AE denotes albedo estimation. According to the quantitative comparison of MAE and RMSE in Table 2, depth maps obtained by DE with AE exhibit higher accuracy than those obtained by existing methods and DE without AE. This can be especially observed in Fig. 5 (a) and Fig. 5 (e).

B. REAL EXPERIMENT

1) REAL EXPERIMENTAL SETTING

In the real experiment, the objects were settled in a dark environment, where the self-contained light of the camera was the only light source. Accordingly, the endoscope camera used the point light sources surrounding the camera lens as the light source. Additional images were captured via smartphone.

2) REAL EXPERIMENT RESULTS

Figure 6 presents original images captured with the smartphone and endoscope camera, along with corresponding esti-

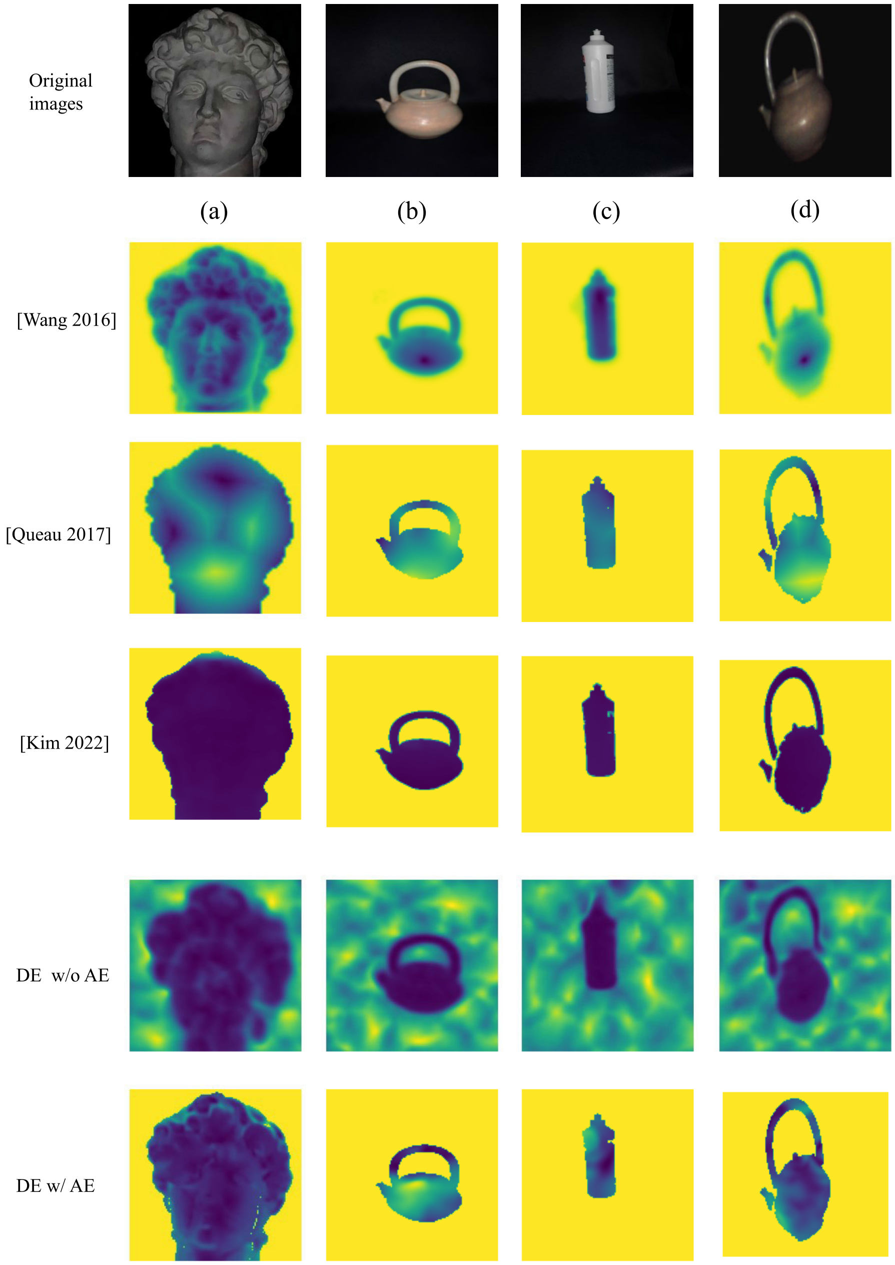


FIGURE 7. Comparison of depth estimation.(a), (b), and (c) were captured using a smartphone. (d) was captured with an endoscope camera. Depth estimation without albedo estimation, depth estimation with albedo estimation and previous work are compared.

TABLE 2. MAE and RMSE comparison of the proposed method and previous work for six images.

Images	Wang [4]		Queau [36]		GLPDepth [37]		DE w/o AE		DE w/ AE	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Fig. 5 (a)	0.2755	0.3057	0.2467	0.2867	0.2658	0.3000	0.2431	0.2757	**0.1176**	**0.1439**
Fig. 5 (b)	0.3233	0.3397	0.2439	0.2546	0.3016	0.3208	0.2941	0.3097	0.1883	0.1977
Fig. 5 (c)	0.2328	0.2600	0.3218	0.3617	0.2613	0.2898	0.2230	0.2536	0.1598	0.1811
Fig. 5 (d)	0.2829	0.3106	0.2649	0.2917	0.2957	0.3276	0.2659	0.2960	0.1419	0.1859
Fig. 5 (e)	0.2565	0.2712	0.2981	0.3290	0.1820	0.2033	0.2314	0.2476	**0.1394**	**0.1573**
Fig. 5 (f)	0.1873	0.2143	0.2786	0.3054	0.2414	0.2674	0.1764	0.2033	0.1594	0.1953

mated albedo and depth maps. Figure 7 depicts a comparison of experimental results between previous work and the proposed method.

In Figure 7, it is apparent that the estimation results obtained by the proposed method are more accurate than those obtained by other methods. For the proposed depth estimation method without albedo estimation, significant amounts of noise appear in the surrounding pixels of the objects. For pixels with an intensity near 0, this method cannot calculate the solution of the PDE with high accuracy.

V. DISCUSSION

In Wang's method, the Newton method is employed for iteration, yielding satisfactory performance for the surrounding pixels of objects. Their method itself is based on the 2D numerical Hamiltonian and fixed-point iterative sweeping method. However, the approach depends on certain artificial viscosities that may influence the solution to the PDE. Moreover, owing to the fixed-point iteration, every pixel's value is dependent upon those of its neighboring pixels, which may explain why the edges of objects appear blurred in the depth maps. Queau's method handles on both orthographic and perspective projection, and uses the ADMM algorithm to solve the PDE. Because it focuses on natural illumination, this method exhibits suboptimal results in dark environment. GLPDepth uses a mix transformer as an encoder and with a light-weight decoder to estimate monocular depth from a single image. The model is trained on the NYU Depth V2 dataset [38]. This end-to-end method also performs well in the dark environment.

In the method of depth estimation without albedo estimation, a mask is not utilized to cover the pixels of surrounding objects. For any surrounding pixels with an intensity near 0, the use of SIREN to solve the PDE cannot yield a sufficiently precise estimate. Consequently, significant amounts of noise appear in the pixels surrounding the object. On the other hand, the ADAM optimizer, which has proven to be a very effective optimizer in machine learning [32], exhibited fast and robust convergence in our study. This may explain why the method of depth estimation without albedo estimation obtained a higher accuracy in depth maps than Wang's method.

For the proposed method of depth estimation with albedo estimation, a mask is utilized to cover the surrounding pixels of objects while minimizing the loss function. Moreover, the albedo parameter ρ is estimated accurately prior to being input to the SIREN network. Without albedo estimation, as is also the case in Wang's method, ρ is assigned arbitrarily. According to the quantitative comparison in Table 2, the depth maps estimated by the proposed method exhibit superior accuracy to those obtained by previous studies.

In terms of computation complexity, our proposed method calculates 396.8k FLOPs, whereas GLPDepth [37] calculates 124M FLOPs. Queau's et al. method [36] calculates approximately 1k FLOPs per iteration per pixel, whereas Wang's and Cheng method [4] calculates 300 FLOPs. Although our proposed method is inferior to certain existing methods in terms of computational complexity, the runtime speed is not an issue in practice due to GPU parallelization.

VI. CONCLUSION

In this study, a depth estimation approach that combines a well-posed SfS problem with albedo estimation has been proposed and implemented. Specifically, the ill-posed problem is converted to a well-posed problem, and learning-based intrinsic image decomposition is conducted to estimate the albedo. Finally, implicit neural representations are used to solve the image irradiance equation. The experiments verify the effectiveness of our proposed method.

Our model does exhibit some limitations. Although materials that are diffuse reflection often yield reasonable approximations, they may cause problems for objects with specularities. Furthermore, our dark experimental environment was set up so that only the point light source of the camera provided lighting. We ignore other possible illumination issues, such as global illumination.

To address the two aforementioned issues, we plan to leverage data-driven methods in future work.

REFERENCES

[1] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on deep neural networks in speech and vision systems," *Neurocomputing*, vol. 417, pp. 302–321, Dec. 2020.

[2] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, May 2021.

[3] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille, "The bas-relief ambiguity," *Int. J. Comput. Vis.*, vol. 35, no. 1, pp. 33–44, 1999.

[4] G. Wang and J. Cheng, "Three-dimensional reconstruction of hybrid surfaces using perspective shape from shading," *Optik, Int. J. Light Electron Opt.*, vol. 127, no. 19, pp. 7740–7751, Oct. 2016.

[5] K. M. Lee and C.-C.-J. Kuo, "Shape from shading with a generalized reflectance map model," *Comput. Vis. Image Understand.*, vol. 67, no. 2, pp. 143–160, Aug. 1997.

[6] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. B. Tennenbaum, "Self-supervised intrinsic image decomposition," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[7] Y. Liu, Y. Li, S. You, and F. Lu, "Unsupervised learning for intrinsic image decomposition from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3245–3254.

[8] W. Bao, R. Komatsu, A. Yamashita, and H. Asama, "Solving well-posed shape from shading problem using implicit neural representations," in *Proc. 2nd Int. Conf. Image Process. Robot. (ICIPRob)*, Mar. 2022, pp. 1–6.

[9] R. Zhang, P. S. Tsai, C. E. Cryer, and M. Shah, "Shape-from-shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug. 1999.

[10] B. K. P. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," Ph.D. thesis, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 1970.

[11] T. Okatani and K. Deguchi, "Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center," *Comput. Vis. Image Understand.*, vol. 66, no. 2, pp. 119–131, 1997.

[12] O. Ikeda, "Shape distortion analysis of the shape-from-shading algorithm using Jacobi iterative method," in *Proc. 1st Int. Symp. 3D Data Process. Vis. Transmiss.*, 2002, pp. 396–399.

[13] B. T. Phong, "Illumination for computer generated pictures," *Commun. ACM*, vol. 18, no. 6, pp. 311–317, Jun. 1975.

[14] A. Ahmed and A. Farag, "Shape from shading for hybrid surfaces," in *Proc. IEEE Int. Conf. Image Process.*, Sep./Oct. 2007, pp. 525–528.

[15] J. Fan, M. Chen, J. Mo, S. Wang, and Q. Liang, "Variational formulation of a hybrid perspective shape from shading model," *Vis. Comput.*, vol. 38, no. 4, pp. 1469–1482, Apr. 2022.

[16] Z. Cao, Y. Wang, W. Zheng, L. Yin, Y. Tang, W. Miao, S. Liu, and B. Yang, "The algorithm of stereo vision and shape from shading based on endoscope imaging," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103658.

[17] E. Garces, C. Rodriguez-Pardo, D. Casas, and J. Lopez-Moreno, "A survey on intrinsic images: Delving deep into Lambert and beyond," *Int. J. Comput. Vis.*, vol. 130, no. 3, pp. 836–868, Mar. 2022.

[18] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–12, 2014.

[19] B. Kovacs, S. Bell, N. Snavely, and K. Bala, "Shading annotations in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6998–7007.

[20] T. Narihira, M. Maire, and S. X. Yu, "Learning lightness from human judgement on relative reflectance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2965–2973.

[21] Z. Li and N. Snavely, "CGIntrinsics: Better intrinsic image decomposition through physically-based rendering," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 371–387.

[22] T. Narihira, M. Maire, and S. X. Yu, "Direct intrinsics: Learning albedo-shading decomposition by convolutional regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, p. 2992.

[23] J. Luo, Z. Huang, Y. Li, X. Zhou, G. Zhang, and H. Bao, "NIID-Net: Adapting surface normal knowledge for intrinsic image decomposition in indoor scenes," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 12, pp. 3434–3445, Dec. 2020.

[24] Y. Ma, X. Jiang, Z. Xia, M. Gabbouj, and X. Feng, "CasQNet: Intrinsic image decomposition based on cascaded quotient network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2661–2674, Jul. 2021.

[25] A. S. Baslamisli, T. T. Groenestege, P. Das, H.-A. Le, S. Kalaoglu, and T. Gevers, "Joint learning of intrinsic images and semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–302.

[26] Y. Yu and W. A. P. Smith, "InverseRenderNet: Learning single image inverse rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3155–3164.

[27] J. Sirignano and K. Spiliopoulos, "DGM: A deep learning algorithm for solving partial differential equations," *J. Comput. Phys.*, vol. 375, pp. 1339–1364, Dec. 2018.

[28] E. Weinan and B. Yu, "The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems," *Commun. Math. Statist.*, vol. 6, no. 1, pp. 1–12, 2018.

[29] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural presentations with periodic activation functions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7462–7473.

[30] E. Prados and O. Faugeras, "Shape from shading: A well-posed problem?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 870–877.

[31] M. Oren and S. K. Nayar, "Generalization of the Lambertian model and implications for machine vision," *Int. J. Comput. Vis.*, vol. 14, no. 3, pp. 227–251, Apr. 1995.

[32] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[33] *ShapeNet*. Accessed: Jan. 11, 2022. [Online]. Available: https://shapenet.org/

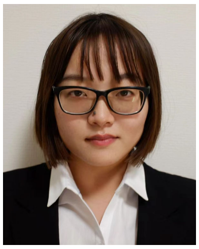
[34] *Blender*. Accessed: Jun. 25, 2022. [Online]. Available: https://www.blender.org/

[35] C. J. Willmott, "Some comments on the evaluation of model performance," *Bull. Amer. Meteorol. Soc.*, vol. 63, no. 11, pp. 1309–1313, Nov. 1982.

[36] Y. Queau, J. Melou, F. Castan, D. Cremers, and J. D. Durou, "A variational approach to shape-from-shading under natural illumination," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2017, pp. 342–357.

[37] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, and J. Kim, "Global-local path networks for monocular depth estimation with vertical CutDepth," 2022, *arXiv:2201.07436*.

[38] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.



WANXIN BAO received the B.S. degree in engineering from the Dalian University of Technology, China, in 2020, and the M.S. degree in engineering from The University of Tokyo (UTokyo), in 2022. Her research interests include computer vision and deep learning.

REN KOMATSU (Member, IEEE) received the B.S. degree in engineering from Yokohama National University, Japan, in 2014, and the M.S. and Ph.D. degrees in engineering from The University of Tokyo (UTokyo), Japan, in 2016 and 2020, respectively. From 2017 to 2018, he was a Visiting Scholar with The Robotics Institute, Carnegie Mellon University, USA. From 2020 to 2022, he was a Project Assistant Professor with UTokyo. Since 2022, he has been an Assistant Professor with UTokyo. His research interests include computer vision, visual SLAM, deep learning, and robot teleoperation.

HAJIME ASAMA (Fellow, IEEE) received the M.S. and Dr.Eng. degrees from The University of Tokyo (UTokyo), in 1984 and 1989, respectively. He was with Riken, Japan, from 1986 to 2002. He became a Professor with the Research into Artifacts, Center for Engineering (RACE), UTokyo, in 2002. He has been a Professor with the School of Engineering, UTokyo, since 2009. He has been the Director of RACE, since 2019. His main research interests include service robotics, distributed autonomous robotic systems, embodied brain science systems, and cognitive ergonomics. He was an AdCom Member of the IEEE Robotics and Automation Society, from 2007 to 2009. He has been a Council Member of the Science Council of Japan, since 2017. He is a fellow of JSME and RSJ. He received the SICE System Integration Division System Integration Award for Academic Achievement, in 2010, and the JSME Award (Technical Achievement), in 2018. He was the Vice President of RSJ, from 2011 to 2012. He has been the President of IFAC, since 2020.

ATSUSHI YAMASHITA (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Faculty of Engineering, The University of Tokyo (UTokyo), in 1996, 1998, and 2001, respectively. From 2001 to 2008, he was an Assistant Professor with Shizuoka University. From 2006 to 2007, he was a Visiting Associate with the California Institute of Technology. From 2008 to 2011, he was an Associate Professor with Shizuoka University. From 2011 to 2022, he was an Associate Professor with UTokyo. Since 2022, he has been a Professor with UTokyo. His research interests include robot vision, image processing, and intelligent robot systems. He is a member of ACM, JSPE, RSJ, IEICE, JSAE, JSCE, JSME, IEEJ, IPSJ, ITE, SICE, and the Society for Serviceology.