IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 11, 2020, accepted June 6, 2020, date of publication June 17, 2020, date of current version June 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003023

Better Together: Shading Cues and Multi-View Stereo for Reconstruction Depth Optimization

ZHE LIANG[1], CHAO XU[1,2,3], JING HU[1,3], YUSHI LI[2,3], AND ZHAOPENG MENG[1]
[1]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3]Shenzhen Graduate School, Peking University, Shenzhen 518055, China

Corresponding author: Chao Xu (xuchao@tju.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFB1701701.

ABSTRACT Passive reconstruction methods such as traditional multi-view stereo are capable to accurate reconstruction results. However, the depth calculation of multi-view stereo encounters significant difficulties, especially when the corresponding points have some degree of inaccuracy or unreliability. In this paper, we make use of geometric and shading cues information in the multi-view stereo approach to construct a robust energy function, which is also effective to optimize the depth information provided by multi-view stereo. This work improves the accuracy of point cloud depth. we evaluated our algorithm in the famous DTU datasets, and established our own dinosaur datasets. All the data is collected by mobile devices under natural light condition, while the reconstruction results are completed and accurate.

INDEX TERMS Depth optimize, MVS, shading cues.

I. INTRODUCTION

The main goal of computer vision is to automatically perceive the world from different dimensions and scales (e.g. 2D images and 3D objects). With the advancement of communication technology, researchers have proposed a large number of valid and efficient 3D reconstruction approaches in recent years. They have been applied in many fields such as autonomous driving, virtual reality, augment reality etc.

As a popular reconstruction method, multi-view stereo [1]–[5] reconstruction recovers 3D information on the basis of the parallax principle which estimates the positional deviation from the corresponding 2D image points. The method is able to achieve qualified reconstruction. Multi-view stereo (MVS) technology is mainly divided into voxel-based [6]–[8], multiple depth map fusion [9], [10] and spatial patch-based approaches [3]. The resolution of the voxel grid limits the accuracy of voxel-based approaches and makes them difficult to handle large-scale scenes. The depth map fusion approaches calculate the depth value of all input images and make these depth value aggregated in the same coordinate system.

Although the stereo method is relatively mature, image patching or regularization is still essential in reducing noise and improving robustness. This results in shape from shading(SFS) [11], which is a method for recovering 3D information from a single image. SFS mimics the human visual system which estimates depth only from a given reflection model that correlates image brightness with local surface normals. Even SFS can produce fine reconstruction results under the premise of regularization, serious problems are caused by textured regions or, equivalently, by non-constant albedo. Because the actual lighting situation is not known, this method is not adapted in the scenes with varying illumination. Therefore, it is necessary to improve the quality of reconstruction by combining the advantages of SFS and MVS.

Many recent works [12]–[15] initialize depth information from MVS and use shadow variations to optimize local depth and capture local details. So we use this framework. Additionally, we improve it by combining geometry and shading-based data terms into a single optimization scheme. In the region where the image gradient changes sharply, greater weight is exerted to alleviate the sharp image gradient. The main contributions of this paper are:

1. The paper presents a framework that combines multi-view stereo and shading cues in order to reconstruct a fine 3D model. The optimization is done with various weighting parameters.

2. This optimization framework enhances local details and removes outliers through specific weight parameters, which balances the accuracy and completeness of reconstruction.

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

II. RELATED WORK

A. MULTI-VIEW STEREO

The multi-view stereo algorithm only relies on images and reconstructs an accurate 3d model based on some reasonable assumptions. For example, the photo-consistency and gradient consistency assumptions are excellent condition. The mainstream local matching models [16], [17] still depend on the local neighborhood which is correspondingly estimated for a given region. These models usually twist the local patch into a common plane or use a matching window of the applicable geometry to prevent serious error. As the pioneer of multi-view stereo, Okutomi and Kanade [1] present the groundbreaking method, which accumulates sum of squared difference (SSD) cost values from different stereo pairs in a set of multiple images and select the depth with the lowest cumulate cost. Since then, many multi-view stereo approaches have been proposed. Furukawa and Ponce [3] approach uses a strategy of matching image patches, relaxing the requirement to find a correspondence for each pixel and filtering the image based on visibility. Tola *et al.* [4] solve the problem of high-resolution image sets by establishing a unique response point between matched pairs of images. Hirschmueller [18] and Galliani *et al.* [19] create multiple nearest neighbor subviews for each view. They calculate the loss value between the two different views and merge it into robust global loss value for optimization. Besse *et al.* [20] balance the photo-consistency constraint and introduce an explicit regularization term. Our work combines the advantages of their researches to optimize each views separately. For the surface representation, we use the surface fitting method proposed by Semerjian [13]. This approach uses bicubic patches to define a surface per view which has continuous depth and normals.

B. SHADING CUES

Combining multi-view geometry and shading cues in 3D reconstruction has been studied for a long time. However, the research of this area is stagnated by the limitation of the experimental environment and hardware. Recently, the method of finding shading cues in multi-view stereo and optimizing has been used by relevant researchers. Wu *et al.* [12] assume that the object is a Lambert surface and only care about general illumination. This work simulates incident lighting through a spherical harmonic function. Jin *et al.* [21] assume a Lambert object with a constant albedo and propose a joint variation method to estimate the shape, normal, and light source. Langguth *et al.* [15] also assume that the Lambert object has a constant albedo and combine the image gradient to optimize the framework based on the Retinex [22] hypothesis. Mauer *et al.* [14] propose a variational method to estimate depth, illumination and albedo by combining shape from shading and stereo. Our shadow cues term is similar to Langguth *et al.* [15], we calculate the albedo in advance based on the Retinex hypothesis, and use a spherical harmonic function to estimate the incident ray.

III. MODEL FRAMEWORK

A. CAMERA SETTING

In this section, we introduce our model framework which integrates multi-view stereo with shading cues. Specifically, we firstly determine our camera settings and the parameterization of the problem, then describe our model framework, including a detailed discussion of all terms. Fig.1 shows the framework of our entire work.

Let us start by defining the underlying camera setting. It consists of n cameras under perspective projection $C_i (i \in 1, \ldots, n-1)$, we use ith image as our example to be the main image located at the origin of the coordinate system. Furthermore, we assume that the projection matrices for all cameras are known as

$$P = K[R|t]. \tag{1}$$

The terms K, R, t represent intrinsic matrix, rotation matrix, translation vector, respectively. $[R|t]$ is the camera's extrinsics matrix, which determines the camera's pose information.

$$K = \begin{bmatrix} f & 0 & o_{ux} \\ 0 & f & o_{uy} \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

denote the corresponding calibration matrix that contains the focal length f and the principal point $o = (o_{ux}, o_{uy})^T$, formalua (3) shows 3D points projection process. We can express the perspective projection $\pi_i : \mathbb{R}^3 \rightarrow \Omega$ of a 3D points $X_v = (X, Y, Z)^T \in \mathbb{R}$ onto the image plane $\Omega \subset \mathbb{R}^2$ of the ith camera as:

$$\pi_i(X_v) = \begin{bmatrix} \dfrac{P_{i,11}X + P_{i,12}Y + P_{i,13}Z + P_{i,14}}{P_{i,31}X + P_{i,32}Y + P_{i,33}Z + P_{i,34}} \\ \dfrac{P_{i,21}X + P_{i,22}Y + P_{i,23}Z + P_{i,24}}{P_{i,31}X + P_{i,32}Y + P_{i,33}Z + P_{i,34}} \end{bmatrix}. \tag{3}$$

By projecting the observed scene, we obtain n images which are denoted by $I_i : \Omega_i \rightarrow \mathbb{R}^3$. Fig.2 shows the specific parameterization and perspective projection. Although there are N cameras, we use two views to express our camera model for simplicity. First we determine a main view I_i and randomly select a two-dimensional point x_v in image I_i, we get the 3D point X_v of x_v by multi-view epipolar geometry. Then, we project the X_v onto j view and can get $P_j(x_v, \mathcal{Z}_i(x_v))$. The formula (4) that converts world coordinates to normalized pixel coordinates is:

$$x = K(RX + t). \tag{4}$$

Therefore, we calculate the coordinate from x_v as following

$$P_j(x_v, \mathcal{Z}_i(x_v)) = K_j(R_j R_i^{-1}(K_i^{-1} x_v \mathcal{Z}_i(x_v) - t_i) + t_j). \tag{5}$$

Here, $i \in N, j \in \mathcal{N}$ and $v \in \mathcal{V}$ denotes index of each image I, index of neighbourhood of I_i and index of each points of I_i, respectively.

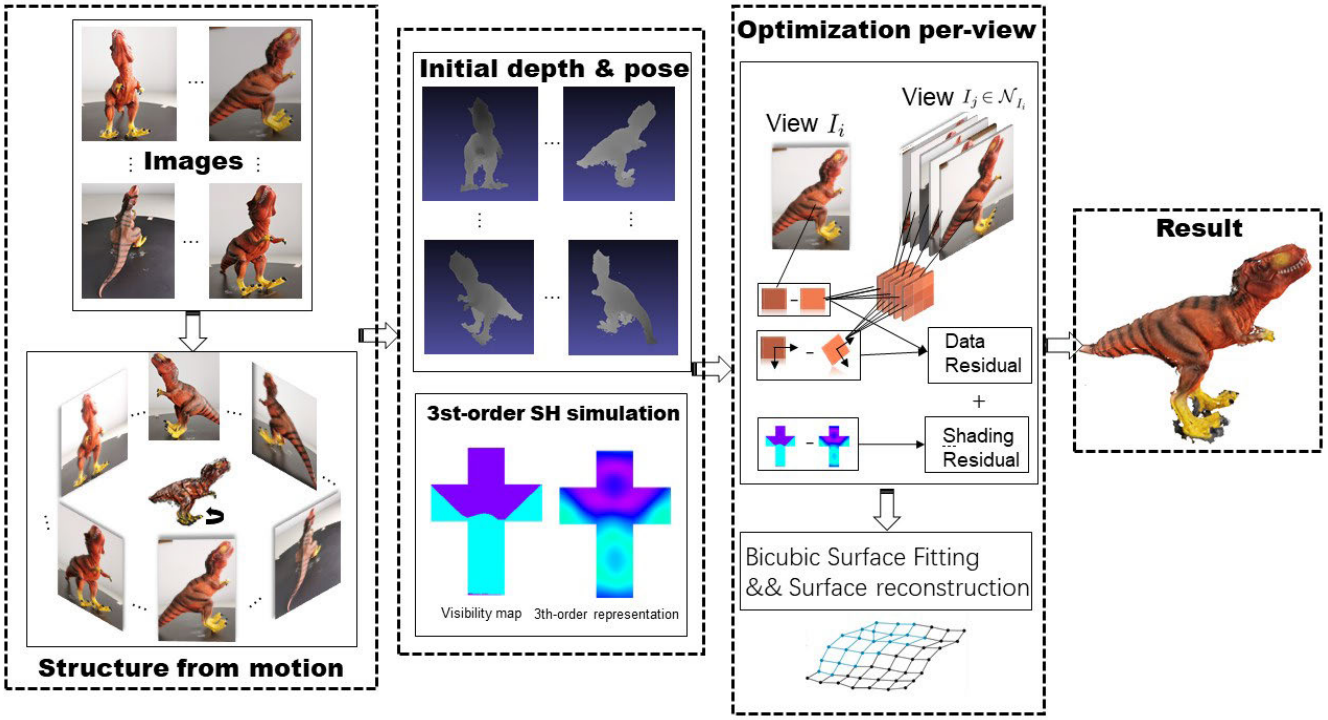


FIGURE 1. Framework:reconstruction depth optimization flow chart. Firstly, the initial depth information and pose information are calculated using structure from motion and semi-global matching algorithm. Secondly, the illumination parameters are estimated and the image is fitted using a spherical harmonic basis function. Our framework calculates the residual of the data item and the shading cues item, and optimizes the depth information by minimizing the residual to get the final model.

B. ENERGY FORMULATION

1) STEREO ENERGY

In multi-view stereo reconstruction, the photo-consistency is a scalar function used to measure the visual compatibility of a three dimensional reconstruction point X_v with a set of images $\mathcal{N}$. A simple photo-consistency function at the 3D points is defined as follows : X_v is projected into each of the visible images, and the similarity of the image texture near their projections is calculated as photo-consistency. In our geometric term, we consider the photo-consistency, i.e. the brightness constancy of projected surface points. Therefore, the photo-consistency energy function is:

$$E_{i,0}(x_v, \mathcal{Z}_i(x_v)) = \frac{1}{n}\sum_{j=1}^{\mathcal{N}} \|I_i(x_v) - I_j(P_j(x_v, \mathcal{Z}_i(x_v)))\|_2^2, \quad (6)$$

where I_i represents the intensity value of the main view, I_j is the intensity value of the adjacent views of I_i. The selection of adjacent views is based on the matched feature number between the main and adjacent views. The gradient consistency is an important assumption for calculating the energy function.

$$E_{i,xy}(x_v, \mathcal{Z}_i(x_v)) = \frac{1}{n}\sum_{j=1}^{\mathcal{N}} \|\mathcal{J}(I_i(x_v)) - \mathcal{J}(I_j(P_j(x_v)))\|_F^2, \quad (7)$$

where $\mathcal{J}(I_i(x_v))$ and $\mathcal{J}(I_j(P_j(x_v, \mathcal{Z}(x_v))))$ are the Jacobian of $I_i(x_v)$ and $I_j(P_j(x_v, \mathcal{Z}_i(x_v)))$, respectively. $||\cdot||_F$ denotes the Frobenius norm. we use a balancing factor γ to give

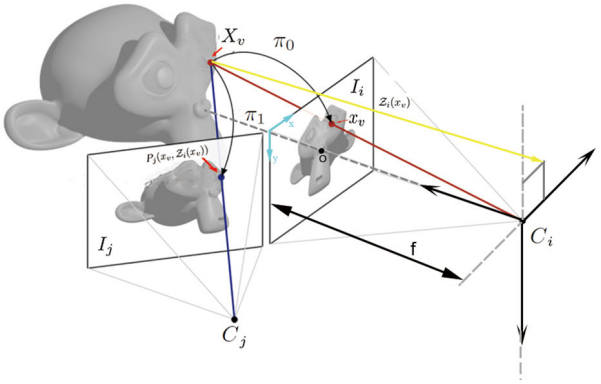


FIGURE 2. Camera parameterization and 3D projection in each image plane. Here we take image I_i as the main view and look for the projection of each point for image I_i in the neighbor view.

different weights to equation (6) and equation (7). Since photo-consistency constraint is accustomed to match the neighboring image set, the robustness is very weak. Because of this, we select the 3×3 patch centered on the projection point in the adjacent image set, and calculate the average light intensity value of the patch to improve robustness.

Finally, in order to get better robustness of assumption w.r.t. outliers and occlusions, we robustify the geometric energy function by applying an auxiliary penalty term [14]:

$$\Phi_{g,0}(s^2) = \Phi_{g,xy}(s^2) = \sqrt{s^2 + \epsilon^2}, \quad (8)$$

where $\epsilon > 0$ is a small constant to ensure differentiability. In summary, we get the energy function of the final stereo

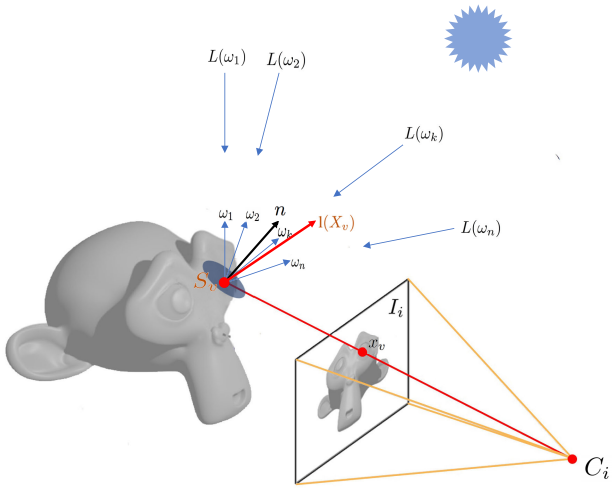


FIGURE 3. The shading cues information graph of image I_i and the image I_i formation process.

data item:

$$E_{i,g}(x_v, \mathcal{Z}_i) = \int_{\Omega_i} (1-\gamma)\cdot\Phi_{g,0}E_{i,0} + \gamma\cdot\Phi_{g,xy}E_{i,xy}dx_i. \quad (9)$$

Here, the choice of γ is to better control the weight of $E_{i,0}$ and $E_{i,xy}$. The experiment demonstrates that this approach is robust and has a positive impact on the results.

2) SHADING CUES ENERGY

Wu *et al.* [12] propose an effective irradiance environment mapping model based on the Legendre polynomial principle. According to the principle of human vision, the natural environment light is mapped into a spherical harmonic model. An effective lighting mode is the basis of reconstructing 3D shapes from images. Similar to the work in [8] which assumes the reconstruction object is lambert surface and the diffuse reflection intensity changes slowly with the direction of the normal vector. The irradiance map and the variation characteristics are described by the quantitative formula (10). The irradiance $\mathbb{E}(\mathbf{n})$ is a function of the surface normal vector $\mathbf{n}$ of the object. This function obtains the analytical expression of the illumination irradiance by integrating the illumination energy of the upper hemisphere:

$$\mathbb{E}(\mathbf{n}) = \int_{\Omega_i} L(\omega)\cdot(\omega^T\cdot\mathbf{n})d\omega, \quad (10)$$

where ω is the negative incident light direction, $\mathbf{n}$ is the unit surface normal, $L(\omega)$ stands for the incident radiance. By multiplying the albedo of the object by the irradiance of the illumination, the irradiance (the gray value) $R_i(x_v)$ of the surface of the object is obtained :

$$R_i(x_v) = \int_{\Omega_i} \rho(x_v)\cdot\mathbb{E}(\mathbf{n})d\omega, \quad (11)$$

where $\rho(x_v)$ is the albedo at x_v. Langguth *et al.* [15] also use the irradiance of Lambert surface to estimate the illumination parameters. The spherical harmonic function is an effective tool to process such signals:

$$R_i(x_v) = \rho(x_v)\cdot\sum_{m=1}^{b^2} l_m H_m(\tilde{\mathbf{n}}(x_v)), \quad (12)$$

where H_m are the spherical harmonics basis functions, and l_m are the corresponding spherical harmonics represented in the spherical harmonics basis, $b-1$ denotes spherical harmonics basis's order. We must solve for the $\rho(x_v)$, H_m, l_m and the unit normal vector $\mathbf{n}$ of x_v. The spherical harmonics basis function H_m are paramatrized of a unit normal vector $\mathbf{n}$ of x_v, and are defined as:

$$\begin{aligned}
H_0 &= 1.0 & H_1 &= n_y \\
H_2 &= n_z & H_3 &= n_x \\
H_4 &= n_x n_y & H_5 &= n_y n_z \\
H_6 &= 2n_z^2 - n_x^2 - n_y^2 & H_7 &= n_z n_x \\
H_8 &= n_x n_x - n_y n_y & H_9 &= (3n_x^2 - n_y^2)n_y \\
H_{10} &= n_x n_y n_z & H_{11} &= (4n_z^2 - n_x^2 - n_y^2)n_y \\
H_{12} &= (2n_z^2 - 3n_x^2 - 3n_y^2) & H_{13} &= (4n_z^2 - n_x^2 - n_y^2)n_x \\
H_{14} &= (n_x^2 - n_y^2)n_z & H_{15} &= (n_x^2 - 3n_y^2)n_x
\end{aligned} \quad (13)$$

We use a third-order spherical harmonic function, resulting in 16 lighting coefficients and 16 spherical harmonic basis functions. The calculation of lighting parameters l_m and albedo $\rho(x_v)$ uses the idea of Langguth *et al.* [15]. l_m is computed ahead of surface optimization using the coarse initial surface model derived from basic stereo. Then albedo $\rho(x_v)$ of x_v is calculated by Retinex-based assumption. The main advantage of this method is that the albedo is fixed as a constant term and the albedo is separated.

In order to separate the albedo from the irradiance function, we take the logarithm of the function and estimate its gradient

$$\nabla\log(R_i(x_v)) = \frac{\nabla\rho(x_v)}{\rho(x_v)} + \frac{\sum_{m=1}^{b^2} l_m\nabla H_m(\tilde{\mathbf{n}}(x_v))}{\sum_{m=1}^{b^2} l_m H_m(\tilde{\mathbf{n}}(x_v))}. \quad (14)$$

According to the Retinex-based assumption: small gradients are caused solely by lighting. Therefore, formula(14) can be changed as:

$$\nabla\log(R_i(x_v)) = \frac{\sum_{m=1}^{b^2} l_m\nabla H_m(\tilde{\mathbf{n}}(x_v))}{\sum_{m=1}^{b^2} l_m H_m(\tilde{\mathbf{n}}(x_v))}. \quad (15)$$

Here, we consider the final shading energy function $E_{i,s}(x_v, \mathcal{Z}_i)$, $\nabla\log(I_i(x_v)) - \nabla\log(R_i(x_v))$ is the residuals generated by shading energy function. So we get our shading cues energy function:

$$E_{i,s}(x_v, \mathcal{Z}_i) = \frac{\nabla I_i(x_v)}{I_i(x_v)} - \frac{\sum_{m=1}^{b^2} l_m\nabla H_m(\tilde{\mathbf{n}}(x_v))}{\sum_{m=1}^{b^2} l_m H_m(\tilde{\mathbf{n}}(x_v))}. \quad (16)$$

Since the albedo is locally constant, it can be seen from the above formula that the irradiance function is only determined by the normal and lighting coefficients of the surface.

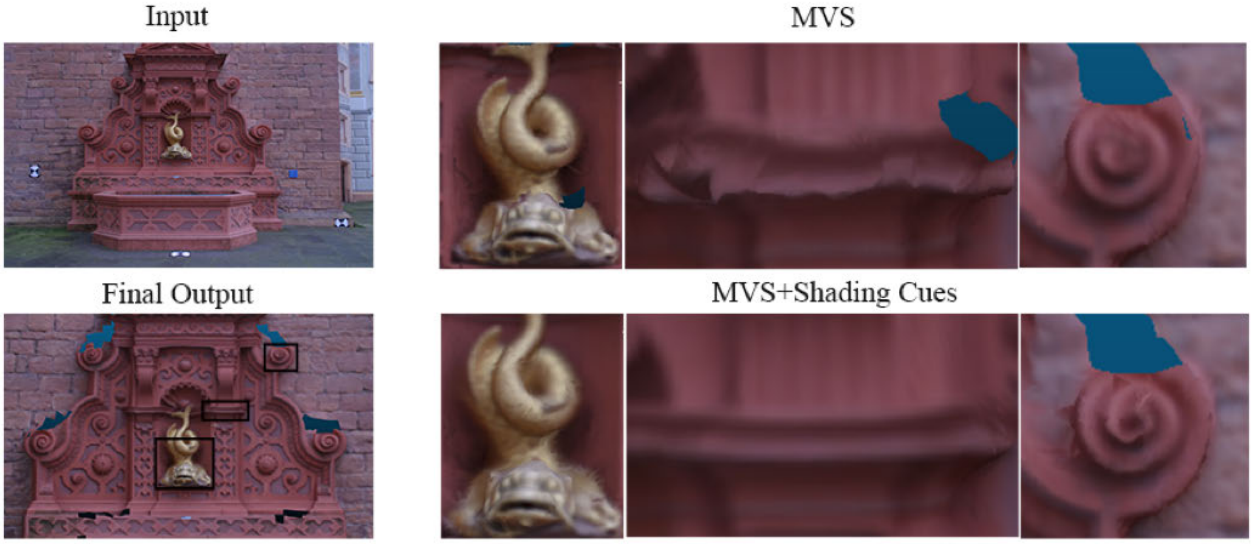


FIGURE 4. The reconstruction result of the *fountain – p11* dataset from Strecha *et al.* [23] The top left if the original input image. The bottom left is the the final output result model. The top right is the local output model only using the MVS method. The bottom right is the local result model obtained by combining MVS and shading cues.

3) JOINT ENERGY

Many works combine different method to build a model framework. A similar approach is used in our work.

The Retinex-based assumption contains an implicit regularization that the albedo is constant in local and can be pre-computed. So we assemble our final energy function:

$$E(x_v, \mathcal{Z}_i) = \sum_{v \in V_i} \alpha E_{i,g}(x_v, \mathcal{Z}_i) + \frac{1-\alpha}{||\nabla I(x_v)||_2^2} \cdot E_{i,s}(x_v, \mathcal{Z}_i), \tag{17}$$

where $E_{i,g}(x_v, \mathcal{Z}_i)$ is stereo geometry data term and $E_{i,s}(x_v, \mathcal{Z}_i)$ is shading cues data term. Our energy function optimizes each pixel in each view.

C. FRAMEWORK OPTIMIZATION

Despite the stereo term and the shading cues term are different, we still use the framework of Semerjian [13] to minimize the residuals generated by our energy function. Semerjian [13] use a bicubic surface representation that relies on computer graphics theory. The shape of the surface depends mainly on surface depth values, the first-order derivative of the depth value along the x-direction, the depth value of the first-order derivative along the y-direction and the mixed second-order derivative of the depth value. It provides a continuous definition of depth values and surface normals that can be applied to the framework to minimize the energy function. After determining the surface representation, we optimize our framework. According to the optimization method proposed by Zollhofer *et al.* [8], we establish a nonlinear least squares problem:

$$E(\mathbf{x}, \mathcal{Z}_i) = ||\mathbf{F}(\mathbf{x})||_2^2, \tag{18}$$

$$||\mathbf{F}(\mathbf{x})||_2^2 = [f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x}), f_4(\mathbf{x})]^T, \tag{19}$$

where $E(\mathbf{x}, \mathcal{Z}_i)$ is the final energy function, and $||\mathbf{F}(\mathbf{x})||_2^2$ is the vector of residuals generated by our energy. $f1$, $f2$, $f3$, and $f4$ represent the residuals of the surface depth value, the first-order derivative of the depth value along the x-direction, the depth value of the first-order derivative along the y-direction and the mixed second-order derivative of the depth value, respectively. The optimized parameters x^* are obtained by solving the minimization problem:

$$\mathbf{x}^* = \underset{\mathbf{x}}{argmin}\, ||\mathbf{F}(\mathbf{x})||_2^2. \tag{20}$$

To this end, we linearize the vector field $\mathbf{F}(\mathbf{x})$ around x_k using a first-order Taylor expansion to obtain an approximation of $\mathbf{F}(\mathbf{x}_{k+1})$:

$$\mathbf{F}(\mathbf{x}_{k+1}) \approx \mathbf{F}(\mathbf{x}_k) + \mathbf{J}(\mathbf{x}_k)\delta_k, \quad \delta_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \tag{21}$$

where $\mathbf{J}(\mathbf{x}_k)$ is the Jacobian matrix of $\mathbf{F}$ evaluated at $\mathbf{x}_k$. We use this approach to transform the nonlinear least squares problem into a linear minimization problem:

$$\delta_k^* = \underset{\delta_k}{argmin}\, ||\mathbf{F}(\mathbf{x}_k) + \mathbf{J}(\mathbf{x}_k)\delta_k||^2. \tag{22}$$

This linear system can be solved by a variety of methods. In the paper, we choose this equation by Gauss-Newton method to find the best least squares solution δ_k^* :

$$\mathbf{J}(\mathbf{x}_k)^T \mathbf{J}(\mathbf{x}_k)\delta_k^* = -\mathbf{J}(\mathbf{x}_k)^T \mathbf{F}(\mathbf{x}_k), \tag{23}$$

where δ_k^* is the update,$\mathbf{J}(\mathbf{x}_k)^T \mathbf{J}(\mathbf{x}_k)$ approximately equal to the Hessian matrix of the optimization parameters. To solve this equation, we first initialize the x_0 using pre-calculated depth values, first-order partial derivatives, second-order mixed derivatives, and illumination parameters, and successively compute the update δ_k^* from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$. In order to solve for the linear update δ_k^*, we choose a preconditioned conjugate gradient(PCG) solver to optimize all

FIGURE 5. Visualization results of scans 1, 6, 11,93 and 106 of DTU dataset [24].Various research work of Multi-View Stereo is showed, such as Camp [5], SMVS [15], PMVS [3], OpenMVS [25], [26] and us.

unknown parameters. In addition, in the optimization strategy, Zollhofer *et al.* [8] propose a coarse-to-fine optimization strategy, which effectively reduces the number of iterations and improves the convergence. First optimize δ_k^* in the coarsest resolution, and then transfer to sub-resolution optimization and finally until the finest resolution δ_k^* reaches the convergence completion iteration and then update the depth information.

IV. EXPERIMENTS

A. EXPERIMENTAL DETAILS

We test our method by combining MVS and shading cues using a variety of datasets. In experiments we make use of a fixed set of solver related parameters: 10 Gauss-Newton iterations per resolution level, 400 PCG iterations, and PCG tolerance is set to 0.005. In addition,α, γ, ϵ are set to 0.8, 0.8, and 0.001, respectively.

We use a novel surface representation [13] that is a bicubic surface. The depth value on each pixel is implicitly generated according to the bicubic parameter. A large number of depth value will be generated in each view. Similar to other depth map fusion methods [27], we fuse depth maps from different view into a unified coordinate system. In order to improve the accuracy and completeness of the reconstruction, we use the visibility-based fusion algorithm [28] to suppress depth occlusions and violations.

After this, we use the statistical outlier removal (SOR) [29] filter, which can remove obvious outliers. In the experiment, the average distance d from each point to the nearest 30 points is calculated, so that the distance of all points in the point cloud constitutes a Gaussian distribution. In order to effectively eliminate outliers, the mean and variance of the Gaussian distribution are set to be d and 0.8, respectively.

In all experiments, depth and albedo are estimated from the original image. The initial depth information of an image is evaluated to provide a coarse initial surface model in order to pre-estimating the lighting parameters l_m.

FIGURE 6. Quantitative results of scans 1, 6, 11,93 and 106 of DTU dataset. We evaluate all methods using the distance metric (lower is better) [24]. Accuracy is measured as the distance from the MVS reconstruction to the structured light reference, and the completeness is measured from the reference to the MVS reconstruction. The selected threshold is 2mm in the all experiment.

B. EXPERIMENT ANALYSIS

The optimization goal is to better balance the accuracy and completeness of the model. To demonstrate the effectiveness of the joint optimization method, we use the *fountain* $-$ *p*11 dataset from Strecha *et al.* [23] to perform MVS reconstruction and joint optimization reconstruction. Figure 4 shows the results of two optimization methods. The local depth information of MVS reconstruction is confusing, which leads to the unevenness of the object surface after meshing. However, the joint optimization reconstruction method optimizes the local depth information to make the depth more reliable.

We use the DTU benchmark dataset [24] for quantitative evaluation. The DTU dataset contains 80 different objects, each covered by 49-64 images of resolution 1600 $\times$ 1200 pixels. The captured scenes have varying reflectance, texture and geometric properties. The images were captured with different lighting conditions using a robot arm to accurately position the cameras. These five models with the most diffuse lighting are used in experiments to evaluate different work. Figure 5 shows the comparison between our algorithm and other research work.

Since our framework only uses a third-order spherical harmonic function to fit the lighting model, it is difficult to handle complex lighting scenes. The lighting in these models is relatively stable, which basically meets the requirements of diffuse reflection. To verify the robustness of the framework, these selected models have different illumination conditions. *Scan*1 and *Scan*6 belongs to diffuse reflection, and the generated results are stable. *Scan*11 and *Scan*93 have slight specular reflection, which can test the robustness of the algorithm. The texture in the model *Scan*106 is relatively flat, and it is difficult to detect local features.

In the experiment, we compare our framework with the currently well-known MVS research work, such as Camp [5], Tola [4], Furu(PMVS) [3], Langguth(SMVS) [15], and the combination of open source SFM software OpenMVG [25] and dense point cloud reconstruction OpenMVS [26]. Figure 4 and Table 1 show Quantitative results of reconstruction. The four key quantitative indicators include completeness, accuracy, median completeness, and median accuracy. In the quantitative Fig.6, we find that the work of camp is more concerned with the completeness of the model. However, Furu and Ponce [3] and Tola *et al.* [4] are more concerned with the accuracy of the model. Similar to our work method, OpenMVS [26] and SMVS [15] pay more attention to both accuracy and completeness of the model. The overall accuracy and completeness are satisfactory. Our overall mean distance is 0.727mm. Compared with others, our method is more robust in terms of accuracy and completeness.

Finally, we create our own dinosaur dataset. This dataset contains 79 dinosaur images taken under soft lighting. We reconstruct the dinosaur model using our algorithm. The completeness and accuracy of the models reconstructed by our approach are relatively good. We also provide download links for such dataset and related parameter descriptions in the Fig.7.

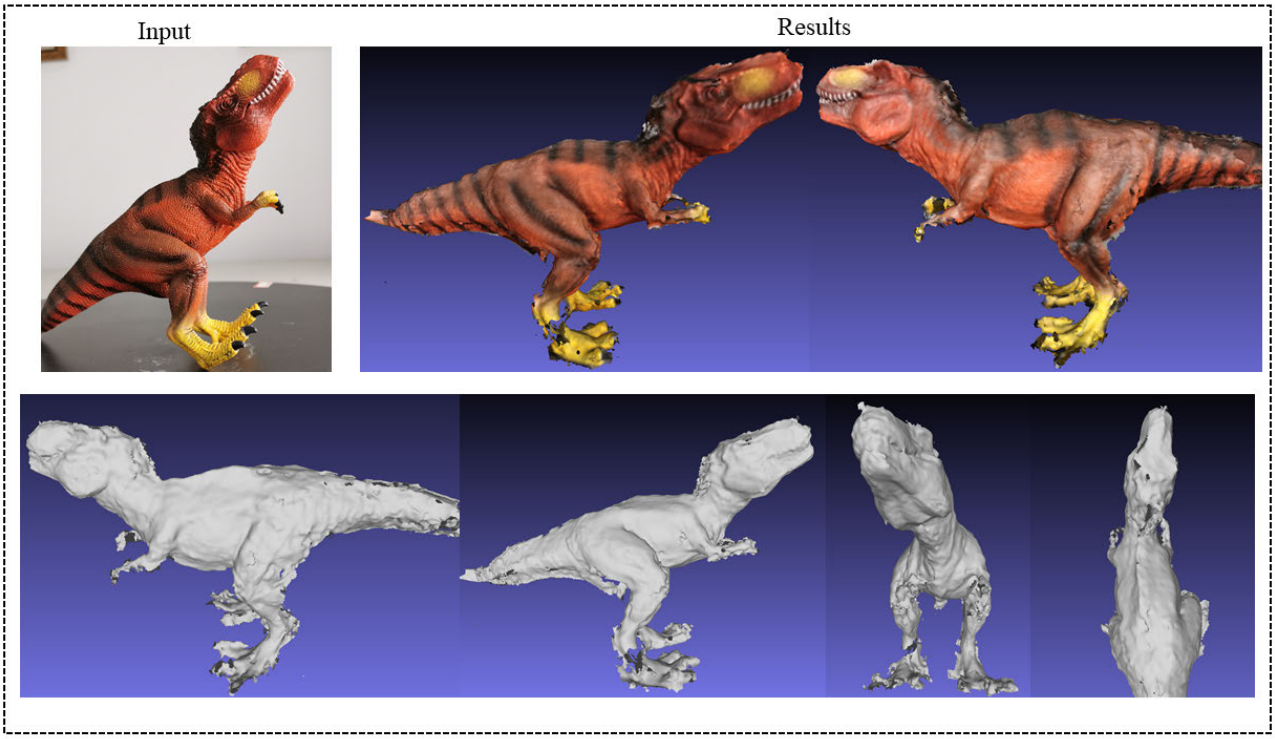


FIGURE 7. Dinosaur datasets. We use the camera module of smartphone to collect images. As can be seen in the figure, we obtain relatively good result. the accuracy and completeness of the model are both relatively high, which indicates that our optimization algorithm is robust. We provide the URL download address of the dataset, please visit https://drive.google.com/open?id=1GVrN_kSkp68bevpvw2e1oa0koME_nGI5.

TABLE 1. Points vs Points (mean distance).

Method Name	Mean Distance(mm)			Median Distance(mm)		
	Acc.	Comp.	Overall	Acc.	Comp.	Overall
Camp[5]	0.695	0.814	0.755	0.446	**0.206**	**0.326**
SMVS[15]	1.046	0.937	0.991	0.691	0.524	0.608
PMVS[3]	0.530	1.352	0.941	0.316	0.579	0.447
OpenMVS[25][26]	0.683	0.815	0.749	0.506	0.588	0.547
Tola [4]	**0.328**	1.374	0.851	**0.215**	0.555	0.385
Ours	0.681	**0.773**	**0.727**	0.503	0.413	0.458

V. CONCLUSION

In the paper we propose a joint method for combining multi-view stereo and shading cues. The method overcomes some shortcomings of multi-view stereo and uses the shading cues of 2D images to refine the model. Our energy function maintains the advantages of these two items and performs depth optimization with various weighting parameters. At the same time, our optimization framework conducts multiple tests on public datasets and our own datasets. Comparing with the previous works, our reconstruction performance performs better in accuracy and completeness. In future work, we will consider the reconstruction algorithms that do not rely on Lambert surface.

ACKNOWLEDGMENT

The authors are grateful to the monographs of these scholars. Without the inspiration and help of these scholars, they will not be able to complete the final writing of this paper.

REFERENCES

[1] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 353–363, Apr. 1993.

[2] P. Labatut, J.-P. Pons, and R. Keriven, "Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[3] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[4] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 903–920, Sep. 2012.

[5] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *Computer Vision—ECCV*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Germany: Springer, 2008, pp. 766–779.

[6] S. Paris, F. X. Sillion, and L. Quan, "A surface reconstruction method using global graph cut optimization," *Int. J. Comput. Vis.*, vol. 66, no. 2, pp. 141–161, Feb. 2006.

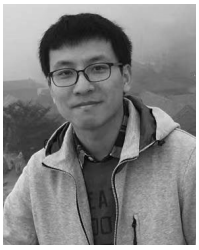
[7] J.-P. Pons, R. Keriven, and O. Faugeras, "Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score," *Int. J. Comput. Vis.*, vol. 72, no. 2, pp. 179–193, Apr. 2007.

[8] M. Zollhöfer, A. Dai, M. Innmann, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner, "Shading-based refinement on volumetric signed distance functions," *ACM Trans. Graph.*, vol. 34, no. 4, p. 96, 2015.
[9] C. Strecha, R. Fransens, and L. Van Gool, "Combined depth and outlier estimation in multi-view stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2394–2401.
[10] Z. Li, K. Wang, D. Meng, and C. Xu, "Multi-view stereo via depth map fusion: A coordinate decent optimization method," *Neurocomputing*, vol. 178, pp. 46–61, Feb. 2016.
[11] B. K. P. Horn, "Height and gradient from shading," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 37–75, Aug. 1990.
[12] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt, "High-quality shape from multi-view stereo and shading under general illumination," in *Proc. CVPR*, Jun. 2011, pp. 969–976.
[13] B. Semerjian, "A new variational framework for multiview surface reconstruction," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 719–734.
[14] D. Maurer, Y. C. Ju, M. Breuß, and A. Bruhn, "Combining shape from shading and stereo: A joint variational method for estimating depth, illumination and albedo," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1342–1366, Dec. 2018.
[15] F. Langguth, K. Sunkavalli, S. Hadap, and M. Goesele, "Shading-aware multi-view stereo," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 469–485.
[16] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch Stereo–Stereo matching with slanted support windows," in *Proc. Brit. Mach. Vis. Conf.*, vol. 11, 2011, pp. 1–11.
[17] N. Einecke and J. Eggert, "Stereo image warping for improved depth estimation of road surfaces," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2013, pp. 189–194.
[18] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 807–814.
[19] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.
[20] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz, "PMBP: PatchMatch belief propagation for correspondence field estimation," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 2–13, Oct. 2014.
[21] H. Jin, D. Cremers, D. Wang, E. Prados, A. Yezzi, and S. Soatto, "3-D reconstruction of shaded objects from multiple images under unknown illumination," *Int. J. Comput. Vis.*, vol. 76, no. 3, pp. 245–256, Mar. 2008.
[22] E. H. Land and J. J. Mccann, "Lightness and retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1970.
[23] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[24] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanaes, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 406–413.
[25] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open multiple view geometry," in *Reproducible Research in Pattern Recognition*, B. Kerautret, M. Colom, and P. Monasse, Eds. Cham, Switzerland: Springer, 2017, pp. 60–74.
[26] D. Cernea. (2020). *OpenMVS: Multi-View Stereo Reconstruction Library*. Accessed: May 10, 2020. [Online]. Available: https://cdcseacave.github.io/openMVS
[27] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixel-wise view selection for unstructured multi-view stereo," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 501–518.
[28] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nister, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[29] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 2011, pp. 1–4.

ZHE LIANG received the bachelor's degree from the School of Physics and Electronic Science, Shanxi Datong University. He is currently pursuing the master's degree with Tianjin University. His research interests mainly include 3D reconstruction and computer vision.

CHAO XU received the Ph.D. degree from the School of Computer Science and Technology, Tianjin University. He is currently a Professor with Tianjin University. His research interests include pattern recognition, affective computing, and knowledge management.

JING HU received the Ph.D. degree from the School of Computer Science and Technology, Tianjin University. She is currently an Assistant Professor with Tianjin University. Her research interests include pattern recognition, machine learning, and knowledge management.



YUSHI LI received the bachelor's degree from the Memorial University of Newfoundland and the master's degree from the City University of Hong Kong. He is currently pursuing the Ph.D. degree with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. His research interests mainly include machine learning, deep learning, computer vision, and computer graphics.

ZHAOPENG MENG received the Ph.D. degree in computer science and technology from Tianjin University. He is currently a Professor with Tianjin University. His research interests include big data computing, Internet of Things software and systems, computer vision and human-computer interaction.