applied sciences

MDPI

Article

FCN-Based 3D Reconstruction with Multi-Source Photometric Stereo

Ruixin Wang [1], Xin Wang [1], Di He [1], Lei Wang [2,*] and Ke Xu [1,*]

[1] Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, Beijing 100083, China; g20189167@xs.ustb.edu.cn (R.W.); s20191301@xs.ustb.edu.cn (X.W.); hedi8888888888@gmail.com (D.H.)

[2] National Engineering Research Center for Advanced Rolling Technology, University of Science and Technology Beijing, Beijing 100083, China

* Correspondence: 2011wanglei2011@gmail.com (L.W.); xuke@ustb.edu.cn (K.X.); Tel.: +86-10-62332598 (L.W.); +86-10-62332159 (K.X.)

Received: 27 March 2020; Accepted: 21 April 2020; Published: 23 April 2020

Abstract: As a classical method widely used in 3D reconstruction tasks, the multi-source Photometric Stereo can obtain more accurate 3D reconstruction results compared with the basic Photometric Stereo, but its complex calibration and solution process reduces the efficiency of this algorithm. In this paper, we propose a multi-source Photometric Stereo 3D reconstruction method based on the fully convolutional network (FCN). We first represent the 3D shape of the object as a depth value corresponding to each pixel as the optimized object. After training in an end-to-end manner, our network can efficiently obtain 3D information on the object surface. In addition, we added two regularization constraints to the general loss function, which can effectively help the network to optimize. Under the same light source configuration, our method can obtain a higher accuracy than the classic multi-source Photometric Stereo. At the same time, our new loss function can help the deep learning method to get a more realistic 3D reconstruction result. We have also used our own real dataset to experimentally verify our method. The experimental results show that our method has a good effect on solving the main problems faced by the classical method.

Keywords: Photometric Stereo (PS); 3D reconstruction; fully convolutional network (FCN)

1. Introduction

Vision-based 3D reconstruction technology can obtain 3D information on the target object from a 2D image in a non-contact manner, which has the advantages of being less affected by the shape of the actual object and giving a more real and robust reconstruction effect. Vision-based reconstruction methods can be roughly divided into active vision methods and passive vision methods. The reconstruction accuracy of active 3D reconstruction methods is relatively high, such as laser scanning and structured light methods, but their cost and complexity are also higher and their reconstruction speed is slow. The passive vision method can make up for the above shortcomings of the active vision method, but still faces challenges in terms of reconstruction accuracy.

As a 3D reconstruction method based on passive vision, the shape from shading (SFS) [1] can analyze the lightness and darkness information in the image and use the reflected illumination model to recover the normal information of the object from a single image. However, a single image contains less information, so the actual reconstruction effect of this method is average. Therefore, in order to improve the shortcomings of the SFS, RJ Woodhan [2] first proposed the Photometric Stereo, using data redundancy to solve the problem of single image reconstruction in SFS due to factors such as shadows and specular reflections, improving the effect and robustness of the reconstruction. On this basis, some

researchers have found that increasing the number of light sources can provide more equations to the solution of unknown parameter parameters [3], thereby compensating for the surface microscopic information missed by the three-dimensional measurement method with three light sources and improving the accuracy of the dimensional measurement, i.e., multi-source Photometric Stereo.

Currently, the research and improvement of Photometric Stereo 3D reconstruction mainly focuses on light source calibration, non-Lambertian reconstruction [4], gradient reconstruction depth [5] and so on. The classic Photometric Stereo method usually assumes that the light intensity on the observation images taken under different illuminations is the same, and the sensor exposure is constant, but these assumptions are difficult to achieve in practical applications. In response to this, Cho et al. [6] developed a method for accurately determining the surface normal direction that is not affected by these factors for situations where the light direction is known but the light intensity is unknown, which improves the accuracy of the Photometric Stereo method in practical applications. Hertzmann et al. [7] proposed a method for calculating the geometry of objects with general reflection characteristics from the image to solve the complex calibration problem of photometric three-dimensional reconstruction, which can be applied to any remote and unknown lighting with almost no calibration operation surroundings.

With the extensive study of deep learning in various fields, neural network frameworks have also been gradually applied to the field of 3D graphics [8,9]. As we all know, the convolutional neural network (CNN) performs well in tasks such as classification and regression. At present, some studies have used CNN to complete three-dimensional tasks. Tang J et al. [10] use the CNN to mix three different three-dimensional shape expressions together, which can bring a better performance to many three-dimensional tasks compared with a single expression. The 3D ShapeNet established by Wu et al. [11] is an earlier proposed 3D reconstruction model of a single image based on voxel representation, using a convolutional depth confidence network to represent geometric 3D graphics as a probability distribution of binary variables on the 3D voxel grid. Its 3D reconstruction was realized by continuously predicting shape types and filling unknown voxels. In a related work, Badrinarayanan et al. [12] established a deep full convolution neural network (FCN) to solve the task of semantics segmentation, which was used to realize the road scene understanding. On the basis of the FCN structure, another network architecture called U-net [13] was established to achieve biomedical image segmentation.

In recent years, the rise of deep learning brings new development direction to the field of machine vision. As a main problem in machine vision, 3D reconstruction has also been widely studied. Eigen et al. [14] adopted a multi-scale deep network with two components, consisting of a coarse-scale network and a fine-grained network, to capture depth information directly. On this basis, a similar neural network architecture was used to process three tasks including depth prediction simultaneously [15], but each task was independently trained by changing its output layer and training objectives. Liu et al. [16] combined the Markov Random Field (MRF) of multi-scale local features and global image features to model the depth of different points and the relationship between them. Other related studies are different from the multi-scale deep network architecture. These include transforming the problem into a classification problem which predicted the likelihood that a pixel would be at any fixed standard depth [17]. Laina et al. [18] used a fully convolutional architecture, encompassing residual learning, to model the ambiguous mapping between monocular images and their corresponding scene depth maps. Xu et al. [19] added a fusion module to the CNN architecture, and the continuous conditional random field (CRF) was used to integrate complementary information on the front-end CNN's multiple side outputs. Li et al. [20] proposed a fast-to-train two-streamed CNN, and the depth and depth gradients were combined either via further convolution layers or directly with an optimization enforcing consistency between the depth and depth gradients. Dechaintre et al. [21] made the result of 3D construction more realistic with a rendering-aware deep network improved by U-net, based on the bidirectional reflectance distribution function (BRDF) [22]. Other related studies include methods based on Bayesian updates and dense [23], the generative adversarial network (GAN) [24], dictionary learning [25], self-augmented convolutional neural networks [26], etc.

For the multi-source Photometric Stereo 3D reconstruction method based on the physical model, using the neural network to simulate the mapping relationship between the real reflection of the object surface and its 3D information is very meaningful research. On the one hand, neural networks can improve the efficiency and accuracy of the multi-source Photometric Stereo, and on the other hand, a lot of existing research on reflection characteristics can also provide a priori knowledge for the neural network algorithms. Although there have been some related studies on learning Photometric Stereo from different perspectives [27–29], the research results in this area are still very limited.

Earlier, Santo H. et al. [29] proposed the use of an FCN in learning Photometric Stereo, and restoring the surface normal of the object from multiple views. After that, Chen G. et al. [28] took the direction of the light source as an input and improved the performance of the algorithm by adding more constraints to the model. Some of the other related studies learnt Photometric Stereo by obtaining the surface normal of the object indirectly. Chen G. et al. [30] proposed a two-stage deep learning structure to solve the uncalibrated Photometric Stereo problem, that is, using a lighting calibration network (LCNet) to recover the light direction and intensity corresponding to the image from any number of images, and then using a normal estimation network (NENet) to predict the normal mapping of the object surface. Compared with the single-stage model, this intermediate supervision effectively reduced the learning difficulty of the network. Moreover, Ikehata S. et al. [31] combined the two-dimensional input image information into an intermediate representation called an observation map to learn Photometric Stereo and used the rotation pseudo-invariance to constrain the network. This method also took the surface normal as the optimization goal. Our method solves the Photometric Stereo 3D reconstruction task from a different perspective. After solving the reflection illumination model, an integration step will be used to restore the three-dimensional topography of the surface, which is also a complicated process. The computational and time cost of this step is also very large, and it may cause cumulative errors and finally cause different degrees of distortion in the reconstructed results. We hope to use depth as the direct optimization goal and obtain the three-dimensional shape of the object surface from end-to-end.

In this paper, we built a U-shaped network structure based on FCN that can obtain the 3D topography of the object surface. By training a parameterized model, we can directly simulate the relationship between physical information such as shadows and reflections on the surface of the object and its depth information. The end-to-end learning can make our method more directly obtain the three-dimensional shape of the object. In addition, we added a regularization constraint on the basis of the general L2 loss function, and the experiments prove that, compared with optimizing the depth value of each pixel directly with the simple L2 loss function [27], this constraint can effectively improve the accuracy of prediction. We also adopted a photometric acquisition setup with a specific configuration to collect a real Photometric Stereo dataset, obtained a high-precision ground truth (GT) using structured light scanning and accurately registered it to the 2D image we collected. The experimental results show that the effectiveness of our method has been verified in a real multi-source Photometric Stereo setup.

The remainder of this study is organized as follows. We first introduce the principle of multi-source Photometric Stereo and the details of our method, including the network structure, our new loss function including two regularization constraints, and the real Photometric Stereo dataset in Section 2. Then the details of our experiments and the experimental results are shown in Section 3. We end with a discussion of our experimental results in Section 4.

2. Materials and Methods

2.1. The Multi-Source Photometric Stereo

The goal of multi-source Photometric Stereo is to recover the original 3D information of the object surface from a set of images with different light source directions. Assume a fixed orthographic camera and directional lighting with multiple equal angle intervals from a fixed latitude line in the upper hemisphere. We assume that a light source from the direction of $\vec{l} \in \mathbb{R}^3$ illuminates a point on

the object surface, and that the surface normal of the point is represented by $\vec{n} \in \mathbb{R}^3$. Then, its pixel intensity can be determined as $I = \vec{\rho} E \cdot \vec{n} \cdot \vec{l}$, where $\vec{\rho}$ is the sensitivity coefficient and E is the light source pre-calibrated brightness, which needs to be obtained through a specific light source calibration method. For the k different light directions $L = [\vec{l}_1, \vec{l}_2, \ldots, \vec{l}_k]^{\mathrm{T}} \in \mathbb{R}^{k \times 3}$, the light intensity can be expressed as $I = [\vec{I}_1, \vec{I}_2, \ldots, \vec{I}_k]^T \in \mathbb{R}^k$, and so the image formation model can be expressed as

$$\begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_k \end{bmatrix} = \vec{\rho} E \cdot \begin{bmatrix} \vec{l}_1 \\ \vec{l}_2 \\ \vdots \\ \vec{l}_k \end{bmatrix} \cdot \vec{n} \,. \tag{1}$$

By solving the above equations, the surface normal direction $\vec{n} = \left(n_x, n_y, n_z\right)$ corresponding to each pixel position will be obtained.

After that, we suppose $(p,\ q, -1) = \left(n_x/n_z,\ n_y/n_z, -1\right)$. Here, p and q are the gradients of a point on the three-dimensional surface in the X and Y directions respectively, and will be formed into the matrices $\vec{P}$ and $\vec{Q}$ in the two-dimensional space. Then, the depth of the 3D surface is defined as $Z(x,\ y)$, and the two gradient values in the directions X and Y are $\triangle Z_x$ and $\triangle Z_y$. Lastly, we use the method of Wu Lun et al. [32] for reference to approximate the actual values $\triangle Z_x$ and $\triangle Z_y$ of the gradient with the $\vec{P}$ and $\vec{Q}$ obtained above. Through the classic two-dimensional integration path algorithm (path integration algorithm, PI), we can obtain a three-dimensional surface with the depth $Z(x,\ y)$.

2.2. Network Architecture

We converted the solution of the mapping relationship from image to depth in the multi-source Photometric Stereo method into an end-to-end optimization process with a large number of parameters. The FCN with encoder-decoder architecture has an outstanding performance in the problem of the pixel-level classification of images; its skip structure combined with the results of different depth layers ensures the long-distance dependence between pixels and the robustness and accuracy of the network and improves the accuracy of the feature extraction. Meanwhile, the network structure of the FCN determines that it can perfectly adapt to any size of input, which is exactly what we needed. Therefore, on the basis of the FCN network structure, we adopted U-net as the basis of our network design.

The architecture of the proposed network is shown in Figure 1. The U-shaped network structure could fully combine the simple features of shallow layer in the decoder stage, so it could also adapt to our small dataset. Our network contained twenty-nine layers, including twenty-one convolution layers, four pooling layers and four up-sampling layers. The activation function of all the convolution operations in the network was ReLU, and we took multiple RGB images from different light source directions containing different degrees of shadow and brightness information as the input of the network. In addition, the network outputted the original RGB images synthesized by the proposed network while outputting the predicted depth—that is, the output of the network was a multi-channel output.

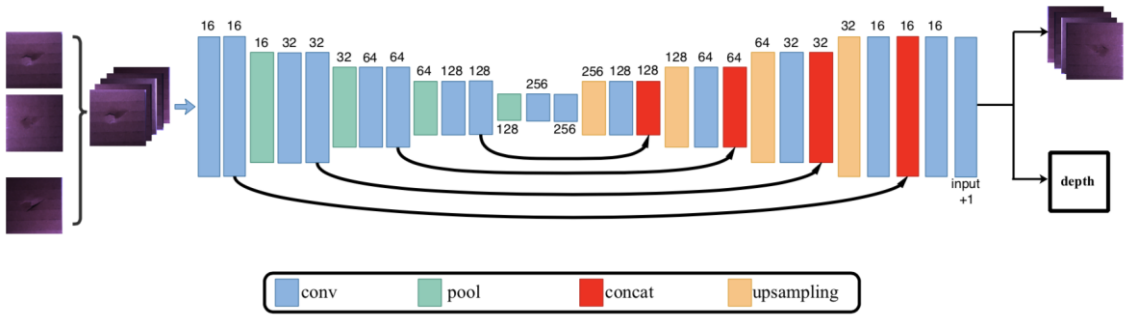


Figure 1. Overview of the proposed network architecture. The network outputted a depth map when given a set of images from the light source direction with different angles as inputs. The kernel sizes for all the convolutional layers were 3×3 and for all up-sampling layers 2×2. Values above the layers indicate the number of feature channels.

2.3. Loss Function

With the U-shaped network based on the code-decode structure, it was easy to lose some details in the training process, and the result of the final output 3D reconstruction was not accurate enough. We propose a loss function which is suitable for the task optimization based on our network structure—that is, we add two regularization constraints on the basis of the L2 loss function, and the training loss for each sample is set to

$$L_{depth} = \|Z - \widetilde{Z}\|^2 + \lambda \|I - \widetilde{I}\|^2, \quad (2)$$

where Z and $\widetilde{Z}$ denote the predicted depth and the ground truth, and respectively, I and $\widetilde{I}$ are the predicted RGB images and the original RGB images. λ is a custom parameter. Here, we have set it to 1×10^4. As described in Section 2.2, our network structure reconstructed the original image of the corresponding light source while predicting the depth value. In the previous experiments, we found that training the network with L2 loss alone can make the network converge, but its reconstruction effect was not good enough. The defect area of the samples had different reflective characteristics under different angles of light, which was an unavoidable phenomenon in the use of the Photometric Stereo method to solve the three-dimensional reconstruction problem. Therefore, the reconstruction results obtained by simply optimizing the depth of each pixel were largely affected by the highlights in the RGB images, and it was not easy to obtain reasonable reconstruction results. Using two regularization constraints, that is, based on the original depth value as the goal of optimization, the original image is also the optimization goal of the network, which could play the role of additional constraints in the network training so as to weaken the influence of the highlight in the input images and make the reconstruction results closer to the real situation. By minimizing the sum of the deviations between the two prediction targets and the GT, our new loss function could improve the effectiveness of feature extraction. Compared with simply predicting the depth value of each pixel position and calculating their loss, this operation, similar to the image restoration, could help correct the prediction results of the network. In Section 4.2, we further evaluate the effectiveness of our new loss function.

2.4. Dataset

2.4.1. The Real-World Dataset

In order to verify the effectiveness of our method, we hoped to use a real-world sample database to train and test our model. At the beginning, we hoped to match our needs to the currently available datasets. However, due to the practical difficulties in 3D data collection, many datasets are based on synthesis or rendering [27,28] and some of them even have no corresponding GT, and so could not be used to train the neural network [33,34]. We think that there are still great differences between

real scene data and rendered simulation data. Therefore, we made a batch of samples by hand and established a real Photometric Stereo experiment platform to collect the dataset we needed.

Our sample database consisted of 100 equal-sized corrugated boards with different degrees of surface damage on them, as shown in Figure 2. The damage on each cardboard was caused by human random. Because the middle part of the corrugated board was partly hollow, the image of the damaged part was very complex under different angles of illumination. The surface features of our samples did not conform to the standard Lambert model, and there were fractures on the surface of the defects which were not a uniform transition. This was not friendly to the classic multi-source Photometric Stereo, as shown in the experimental results.

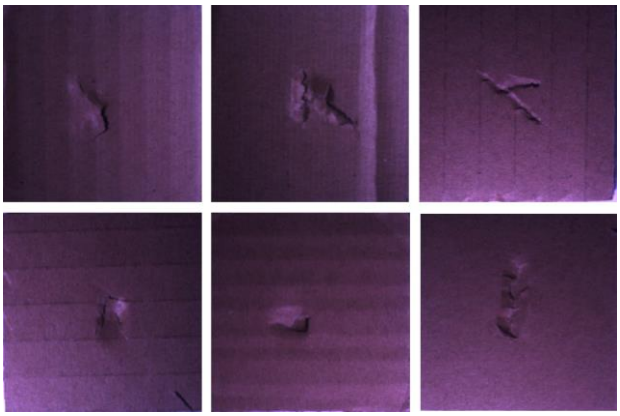
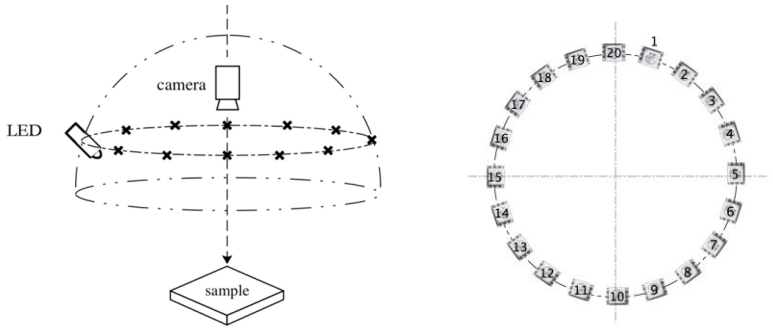


Figure 2. Examples of our real-world dataset.

2.4.2. The Photometric Acquisition Setup

We set up a real photometric stereo experiment platform to collect the images needed for training, as shown in Figure 3a. The camera and the circular light frame were fixed by a frame including clamping devices to ensure that the light conditions of each acquisition were determined and consistent, and the light frame was fixed with the camera (Automation Technology GmbH, Bad Oldesloe, Germany) at its center. The arrangement of the circular light frame is shown in Figure 3b. We designed our light sources as 20 white LED bulbs of the same size (60 degrees) as the scattering angle and fixed them on a circular ring. The angle interval between each adjacent white LED bulb was 18 degrees.



(**a**) (**b**)

Figure 3. (**a**) The photometric acquisition setup of the multi-source Photometric Stereo. (**b**) The light configurations of our proposed setup.

2.4.3. Data Capture

Through the program control, we lit up the LED bulbs in each direction in order and collected 2000 images corresponding to 100 samples in turn, all of which were captured in the dark room. We used 95

samples as the training set and the remaining 5 as the test set. The collected samples were cropped and then resized to the pixel size of 256∗256, which was convenient for network training and better fitting function. We also obtained the GT of each object by line structured light scanning and accurately registered them on the two-dimensional images we collected, as shown in Figure 4.

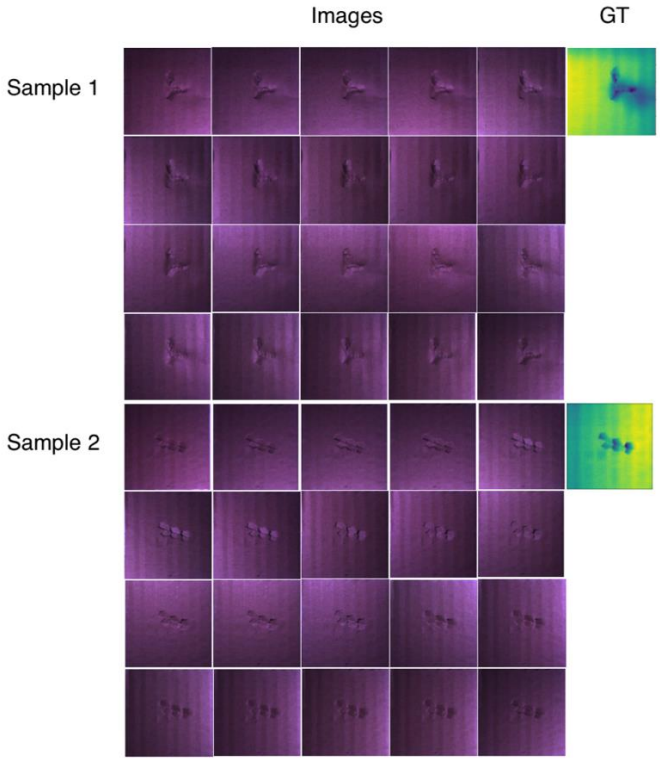


Figure 4. Examples of the collected images and their corresponding ground truth (GT).

3. Results

3.1. Implementation Details

We used a Tensorflow (tensorflow_gpu-1.8.0-cp35-cp35m-win_amd64.whl) with a Nvidia GTX2080 graphics card to implement and train the proposed network. The training process used a batch size of 16 for 100 epochs. The loss function was optimized using the Adagrad Optimizer and the learning rate was 1×10^4. We initialized the weights with a zero-mean Gaussian distribution and a standard deviation of $\sqrt{2/fin}$, where the fin was the number of input units in the weight tensor.

For each sample object, we selected two-dimensional images from the light source direction at 5 equal angle intervals to train our network. That is to say there were 4 kinds of light source combinations for the 20 images collected from each actual sample that could be used as an input for our network. In this way, the size of the training set was 380 (95*4). We used it as a type of data augmentation to train our network. The results predicted by the general loss function optimization network were also evaluated by the same setup. In addition, all 20 images collected for each sample were also used to test the classic multi light source photometric stereo method as a comparative experiment.

3.2. Error Metrics

As shown in Table 1, we used five indices to quantitatively evaluate several methods involved in this experiment which are widely used in the error analysis and accuracy analysis of deep estimation based on deep learning [14,18,27]:

1. root mean squared error(rms)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|d_i - d_i^*\right|^2}, \tag{3}$$

2. average relative error(rel)

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\left|d_i - d_i^*\right|}{d_i^*}, \tag{4}$$

3. threshold accuracy(δ)

$$\begin{gathered}\delta = \frac{1}{N}\sum_i \eta_i, \\ \eta_i = \begin{cases} 1 & if\ T < t \\ 0 & if\ T \geq t \end{cases}, \\ T = max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right), t \in [1.25, 1.56, 1.95], \end{gathered} \tag{5}$$

where d_i^* and d_i are the GT and predicted depths respectively of each pixel, and according to the different values of t, the results of δ(t) are divided into three grades.

Table 1. Quantitative evaluation of our method in comparison to the reference method using the L2 norm. Lower is better for rms and rel and higher is better for δ(t).

Methods	rms	rel	δ(1.25)	δ(1.56)	δ(1.95)
L2 Norm	0.3770	0.3552	0.6492	0.9859	0.9906
Ours	0.2797	0.2473	0.2359	0.9727	0.9937

4. Discussion

4.1. Compared with the Classic Multi-Source PS

In some classic multi-source Photometric Stereo 3D reconstruction methods, the effect of highlights on the results is removed by a selecting method—that is, some images that contain severe highlight reflections will not participate in the calculation. However, this loses a lot of meaningful information contained in the highlight position, even reducing the rationality of the prediction. The characteristics of the neural network determined that it could be biased towards learning information from the input that was more relavent to the correct results. Therefore, using the neural network to learn will not lose the useful information of the highlight position itself, but can also help to reduce inaccurate predictions caused by specular reflections and noise. Furthermore, we represented the optimized target as the depth value of each pixel. Compared with other representations such as point clouds or voxel grids, such 2D representations make the computational cost of our network less.

In order to verify the practical significance of the neural network used to learn Photometric Stereo for 3D reconstruction, we compared the results of the classic multi-source Photometric Stereo method (BASELINE) and ours with GT to conduct a qualitative analysis. We reconstructed the target surface with the BASELINE method using 5, 10 and 20 two-dimensional images taken under the illumination of light sources with equal angle intervals, as shown in Figure 5c–e. The BASELINE method had an obvious effect on the reconstruction of the corrugates which excessive smoothly, but the cast shadow and attached shadow caused by the fracture led to an anomaly in the 3D information extraction at the deeper fractures (Sample 3). However, there was a smooth transition in ours at the fracture site, which made our prediction more reasonable. For the defect surface with more small cracks, ours

could not reproduce all the details perfectly. In comparison, the BASELINE results lost more surface information, and the smooth inclined position with little feature information could not present a reasonable three-dimensional shape. In addition, because the surface features of the target samples did not conform to the standard Lambert model, the reflection around the defect resulted in different degrees of bulge in the transition from the plane to the defect in the reconstruction surface of the BASELINE method (Sample 2). Our method took GT as the direct optimization goal, which could minimize the influence of the highlights in the input on the correct prediction results.

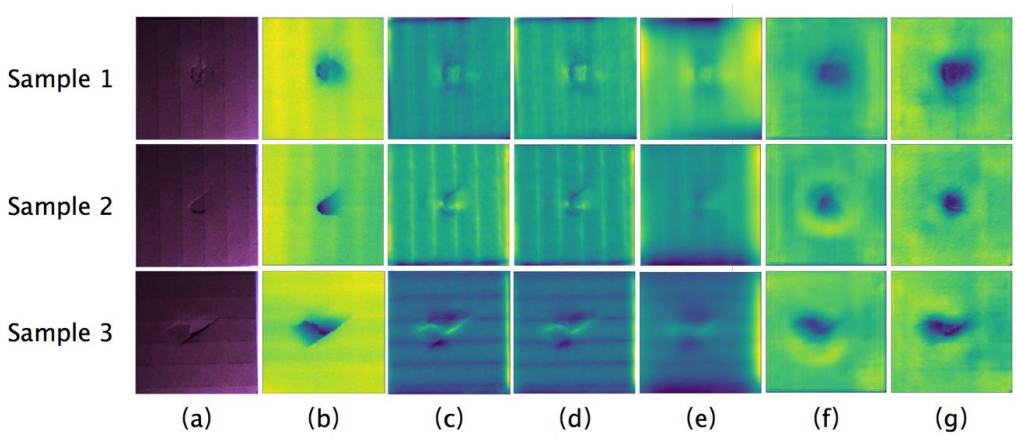


Figure 5. Examples of the results of our method and others. The samples in the example are all from the test set: (**a**) the first column shows the basic 2D information of this sample; (**b**) the corresponding ground truth data (calculated using linear structured light scan) are shown in the second column; (**c**–**e**) the third to fifth columns demonstrate the classic multi-source Photometric Stereo approach using 5,10, and 20 input images; (**f**) the sixth column shows the results using a general L2 loss function; (**g**) the result estimated by our network is shown in final column.

4.2. Effectiveness of the New Regularization Constraints

Most of the recent studies use the normal vector solved by the reflected illumination model as the optimization target, but the solution from the normal vector to the depth is also a complex problem. To verify the effectiveness of our new regularization constraints, we used the proposed network and the same configuration, but used a common loss function with a general L2 norm to train the dataset, which was used in the recent related work [27]. By comparing group (f) and group (g) of these three samples in Figure 5, we can find that ours(g) contained more details than the method with the general L2 norm did (f). Since our optimized target also included the original image of the object, generating the input images could help our network to correct the prediction of the depth, so that the reconstruction result was closer to the real. Thus, ours was clearer for the reconstruction of the simple sample surface (corrugates), and the transition of the cracks on the defects was also smoother (Sample 3). As shown in Figure 5 (Sample 2, Sample 3), there was an abnormal bulge around the defects as we can see in group (f), but from the original image and GT corresponding to the sample, this did not conform to the real situation. However, ours had a good effect on the optimization of this special position—that is, the transition from plane to defect was more reasonable. In addition, Table 1 shows the quantitative analysis results of our method and the general L2 norm. It can be seen from the table that our method significantly improved on the parameters rms and rel. However, the threshold accuracy of ours was slightly lower than that of the L2 norm. The main reason for this, we think, was that the restoration of the images made our reconstruction results closer to reality rather than only taking the depth GT as the optimization standard. Therefore, the accuracy of the depth prediction was lower than that of GT, but it could also get the same level of L2 loss within a certain accuracy range.

5. Conclusions

In this paper, we proposed an effective improvement method aimed at problems such as the complex calibration process and low reconstructing speed faced by the traditional multi-source Photometric Stereo method in 3D reconstruction tasks to improve its accuracy and efficiency. Hereto, we trained the neural network model with a large number of parameters in an end-to-end way to simulate the relationship between physical information, such as shadow and reflection on the surface of the object, and depth information in the multi-source Photometric Stereo. In contrast, our method was superior to the classic algorithm in terms of efficiency and accuracy. In addition, we proposed a new regularization constraint, which improved the effectiveness of feature extraction by minimizing the sum of the loss of the two prediction targets, making the prediction closer to reality.

Author Contributions: Conceptualization, D.H. and K.X.; methodology, R.X. and D.H.; software, X.W. and D.H.; validation, L.W. and K.X.; formal analysis, R.W.; investigation, R.W.; resources, R.W., X.W.; data curation, R.W., X.W.; writing—original draft preparation, R.W.; writing—review and editing, K.X.; visualization, L.W.; supervision, K.X.; project administration, L.W.; funding acquisition, K.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (no.2018YFB0704304), and the National Natural Science Foundation of China (grant number 51674031 and 51874022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Horn, B.K.P. Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View. Ph.D. Thesis, Department of Electrical Engineering, MIT, Cambridge, UK, 1970.
2. Woodham, R.J. Photometric method for determining surface orientation from multiple images. *Opt. Eng.* **1980**, *19*, 139–144. [CrossRef]
3. Xu, K.; Wang, L.; Xiang, J.; Zhou, P. Three-dimensional defect detection method of metal surface based on multi-point light source. *China Sci.* **2017**, *12*, 420–424. (In Chinese)
4. Sun, J.; Smith, M.; Smith, L.; Midha, S.; Bamber, J. Object surface recovery using a multi-light photometric stereo technique for non-Lambertian surfaces subject to shadows and specularities. *Image Vis. Comput.* **2007**, *25*, 1050–1057. [CrossRef]
5. Wang, L.; Xu, K.; Zhou, P.; Yang, C. Photometric stereo fast 3D surface reconstruction algorithm using multi-scale wavelet transform. *J. Comput. -Aided Des. Comput. Graph.* **2017**, *29*, 124–129. (In Chinese)
6. Cho, D.; Matsushita, Y.; Tai, Y.W.; Kweon, I. Photometric Stereo Under Non-uniform Light Intensities and Exposures. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
7. Hertzmann, A.; Seitz, S.M. Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1254–1264. [CrossRef] [PubMed]
8. Li, X.; Dong, Y.; Peers, P.; Tong, X. Synthesizing 3D Shapes from Silhouette Image Collections using Multi-projection Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5535–5544.
9. Sun, C.Y.; Zou, Q.F.; Tong, X.; Liu, Y. Learning Adaptive Hierarchical Cuboid Abstractions of 3D Shape Collections. *ACM Trans. Graph.* **2019**, *38*, 1–13. [CrossRef]
10. Tang, J.; Han, X.; Pan, J.; Jia, K.; Tong, X. A Skeleton-bridged Deep Learning Approach for Generating Meshes of Complex Topologies from Single RGB Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4541–4550.
11. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3D ShapeNets: A Deep Representation for Volumetric Shape Modeling. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
12. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]

13. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
14. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Advances in Neural Information Processing Systems, Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada 8–13 December 2014*; Curran Associates, Inc.: New York, NY, USA, 2014.
15. Eigen, D.; Fergus, R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
16. Liu, F.; Chung, S.; Ng, A.Y. Learning depth from single monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *18*, 1–8.
17. Ladicky, L.; Shi, J.; Pollefeys, M. Pulling Things out of Perspective. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
18. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper Depth Prediction with Fully Convolutional Residual Networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016.
19. Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; Sebe, N. Multi-Scale Continuous CRFs as Sequential Deep Networks for Monocular Depth Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
20. Li, J.; Klein, R.; Yao, A. A Two-Streamed Network for Estimating Fine-Scaled Depth Maps from Single RGB Images. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017.
21. Deschaintre, V.; Aittala, M.; Durand, F.; Drettakis, G. Single-Image SVBRDF Capture with a Rendering-Aware Deep Network. *ACM Trans. Graph.* **2018**, *37*, 1–5. [CrossRef]
22. Nicodemus, F.E. Geometrical Considerations and Nomenclature for Reflectance. *NBS Monogr.* **1977**, *160*, 4.
23. Hermans, A.; Floros, G.; Leibe, B. Dense 3D semantic mapping of indoor scenes from RGB-D images. In Proceedings of the IEEE International Conference on Robotics & Automation 2014, Hong Kong, China, 31 May–7 June 2014.
24. Yoon, Y.; Choe, G.; Kim, N.; Lee, J.Y.; Kweon, I. Fine-scale Surface Normal Estimation using a Single NIR Image. In Proceedings of the European Conference on Computer Vision 2016, Amsterdam, The Netherlands, 11–14 October 2016.
25. Xiong, S.; Zhang, J.; Zheng, J.; Cai, J.; Liu, L. Robust surface reconstruction via dictionary learning. *ACM Trans. Graph.* **2014**, *33*, 1–12. [CrossRef]
26. Li, X.; Dong, Y.; Peers, P.; Tong, X. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graph.* **2017**, *36*, 45. [CrossRef]
27. Liang, L.; Lin, Q.; Yisong, L.; Hengchao, J.; Junyu, D. Three-Dimensional Reconstruction from Single Image Base on Combination of CNN and Multi-Spectral Photometric Stereo. *Sensors* **2018**, *18*, 764.
28. Chen, G. PS-FCN: A Flexible Learning Framework for Photometric Stereo. In Proceedings of the European Conference on Computer Vision (ECCV) 2018, Munich, Germany, 8–14 September 2018.
29. Santo, H.; Samejima, M.; Sugano, Y.; Shi, B.; Matsushita, Y. Deep Photometric Stereo Network. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy, 22–29 October 2017.
30. Chen, G.; Han, K.; Shi, B.; Matsushita, Y.; Wong, K.K. Self-calibrating Deep Photometric Stereo Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 16–20 June 2019.
31. Ikehata, S. CNN-PS: CNN-based Photometric Stereo for General Non-Convex Surfaces. In Proceedings of the European Conference on Computer Vision (ECCV) 2018, Munich, Germany, 8–14 September 2018.
32. Wu, L.; Wang, Y.; Liu, Y. A robust approach based on photometric stereo for surface reconstruction. *Acta Autom. Sin.* **2013**, *39*, 1339–1348. (In Chinese) [CrossRef]

33. Alldrin, N.; Zickler, T.; Kriegman, D. Photometric stereo with non-parametric and spatially-varying reflectance. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
34. Einarsson, P.; Chabert, C.F.; Jones, A.; Ma, W.C.; Lamond, B.; Hawkins, T.; Bolas, M.; Sylwan, S.; Debevec, P. Relighting Human Locomotion with Flowed Reflectance Fields. In *Eurographics Workshop on Rendering, Proceedings of the 17th Eurographics Conference on Rendering Techniques Nicosia, Cyprus, 26–28 June 2006*; Eurographics Association: Goslar, Germany, 2006.