Augmenting Depth Estimation from Deep Convolutional Neural Network using Multi-Spectral Photometric Stereo

Yisong Luo[1], Hengchao Jiao[1], Lin Qi[1],Junyu Dong[1,*], Shu Zhang[2], Hui Yu[2]
[1]*Ocean University of China,Qingdao, China*
[2]*University of Portsmouth, Portsmouth, UK*
dongjunyu@ouc.edu.cn

***Abstract*—Multi-Spectral Photometric Stereo can recover surface normals from a single image, but requires an initial estimate of the normals due to the tangle of the illumination, reflectance and camera responses on each of the RGB channels. Instead of employing a depth sensor or binocular stereo device, in this paper, we propose a method to estimate fine-scale geometry structures with the popular Deep Convolutional Neural Networks (CNNs). We train the network with rendered images of synthetic 3D objects, and apply the trained model with real world data. The CNN is used to estimate a rough prediction of the depth, then the normals from Multi-Spectral Photometric Stereo are progressively refined accordingly. Experiments demonstrate the competitive results of our method for improving the depth estimation.**

***Index Terms*—deep estimation, convolutional neural network, multi-spectral photometric stereo**

1. Introduction

Computer vision aims to simulate the human visual functionalities by computational algorithms, including recognizing, tracking, and measuring objects by cameras and computers. A fundamental problem in computer vision is to study how to obtain information about the structure and properties of the 3D world from the 2D images of the scene since the 2D cameras is still the dominating sensor in this area. Therefore, the depth estimation, which extracts the depth of information automatically and effectively from one or more images, is an key research in the field of computer vision.

Traditionally, there are two different kinds of methods to tackle this problem: (a) Active Vision based ones, such as the shape from shading ([1], [2]) and photometric stereo ([3], [4]); (b) Passive Vision based methods, including multi-view stereo ([5], [6]) and structure from motion ([7], [8]). The Active Vision based ones are achieved with the shading cues caused by the interactions of object's surface and lightings. They focus on the reconstructions at per-pixel level. On the other hand, the Passive Vision based ones estimate the depth according to the projective geometry and photo-grammetry, which are more effective in estimating the depth in a metric way. Commonly, most of these methods need a sequence of images (typically two or more) of a 3D object or scene to acquire the final results. Furthermore, the sequence of images required by Active Vision need to be captured by a stationary camera.

Apart from the methods based on 2D RGB cameras, there are several dedicated equipments that can obtain 3D geometry structure. Those device are achieved by the techniques such as the TOF (Time-of-Flight) based ones ([9]) and the Structure Light based ones. The TOF based ones calculate the travel time of the laser beam between the sensor and the scene for 3D reconstruction, for example, the Laser scanner. It can get fine results in high accuracy, however also with a high cost, which limits its range of application. The Structure Lights based ones such as Microsoft's Kinect and Intel's RealScene, are widely applied in commercial use and scientific and technical research because of low cost. The Kinect analyzes the patterns projected by the infrared projector for 3D reconstruction while RealScene achieve the recovery with the projected infrared grid. A lot of works (e.g. [10]) have been proposed to improve the results of those RGB-D cameras. Though these works have made great progresses on depth estimation, they still require specific system configuration and commonly with heavy computation cost. To this end, we propose a novel 3D reconstruction method using Deep CNNs to get minimize these limitations.

Deep learning has been achieving various breakthroughs in computer vision. Especially, the Deep convolutional neural networks (CNNs) are capable of dealing with many computer vision problems such as object detection, image segmentation, image classification, scenes understanding, and depth estimation. Recently, several approaches [11], [12] have been proposed to estimate the depth using CNNs. Most of them aimed at estimating the depth of a scene such as a indoor living room or a outdoor street. However, the depth estimations obtained by those deep learning based methods are commonly coarse and cannot be used in the circumstances that require high accuracy.

In this paper, we focus on the estimation of the depth for fine-scale objects, rather than a scene. We combine the deep CNN with the mechanism of photometric stereo (PS). In this case, the deep CNN can handle the problem of estimating the depth from one single image. The obtained coarse result can improve the precision for PS. First, we train our network with the rendered images of 3D models

from the ShapeNet dataset. The images are rendered on the fly during the training of the network. After the model is acquired, we estimate the depth map using the pre-trained network for the image of a real object which is similar to synthetic models. Then, this depth estimation is treated as input for the proposed PS method to compute surface normal map of the object. Finally, we convert the depth data into normal data to correct the depth estimation.

2. Related Work

2.1. Photometric Stereo

Photometric stereo is a well studied topic for image based 3D reconstruction technique that can acquire very high quality of reconstruction results [13]. A stationary camera captures a series of images (at least 3) of a 3D object under multiple controlled illuminations. The intensity with the same image coordinate changes across these images with respect to the various directions of illuminations. Accordingly, the surface normals of this object can be computed based on corresponding intensities and lighting direction. The depth information is integrated by normal afterward, and then a fine detailed reconstruction of the object is obtained.

The photometric stereo is first introduced by Woodham [3]. He constrained the method to be effective only with the Lambertian surface reflectance model, which is an hypothesis that assume the albedo of each points on object is constant. Coleman *et al.* [14], Nayer *et al.* [15], Lin *et al.* [16] and Jensen *et al.* [17] relaxed the assumption to non-Lambertian reflectance models such as bidirectional reflectance distribution function (BRDF) [18] and bidirectional surface scattering distribution function (BSSRDF) [17], *etc*. Several works dealt with the frequent presence of shadows and specular in an image (e.g., [19]). However, these methods suffered from the same restriction, which is that all of images must be captured by the camera relative still to the scene while the illumination is changing. It means photometric stereo cannot reconstruct an object in motion.

To relax this restriction, Drew *et al.* [4] and Kontsevich *et al.* [20] initially proposed a multi-spectral photometric stereo technique, which can obtain a detailed geometry structure from a single image. In essence, multi-spectral photometric stereo is photometric stereo with colored light. Unlike photometric stereo which photographs objects under varying white lights and processes gray-scale images, the multi-spectral PS captures a RGB image under three colored light sources at one time. Anderson *et al.* [10] extended the multi-spectral PS to reconstruct objects with continuous chromaticities. Though this further expanded the applications of multi-spectral PS, the depth estimation is still cannot be acquired without the prior knowledge of lighting directions.

The commercial depth sensors such as Kinect and RealScene can obtain 3D information of objects in real time without any prior knowledge about the object or illumination. There are existing works using depth sensors to improve depth estimation results obtained by photometric stereo. For example, Zhang *et al.* [21], and Yu *et al.* [22] introduced several sensor fusion schemes that combine active stereo with photometric stereo. They prevented the quantization effect of Kinect and enhanced surface details. Moreover, their methods worked well under circumstances where illumination varies with the minimum intensity and ambient light conditions. However, these methods highly relied on the result of depth sensors and required high computational costs.

Even though the traditional photometric stereo method can obtain robust results, the reconstruction process is still quite limited when applying the method to estimate the surface depth of objects from a single RGB image practically. The fine results require ideal assumptions, extra system configuration, heavy computation time, and the calibration of lighting directions. To deal with these limitations, we propose a scheme to enhance the traditional photometric stereo by a deep convolutional neural network.

2.2. Machine Learning for Depth Estimation

Machine learning have been achieved many accomplishments in the field of computer vision. Many existing works focused on depth estimation via machine learning approaches. Xiong *et al.* [23] applied dictionary learning to jointly optimize the geometry and connectivity construction. It used triangular meshes for representing the surface of object. However, the original dictionary must be given via a dense point cloud, which means that their method is useful only for refining geometry structures that have been already reconstructed beforehand.

3. Method

Recently, compared with other machine learning methods, the deep convolutional neural networks (CNN) has received more attentions from the researchers in many areas. The methods based on deep CNNs can estimate depth from a single image due to their great abilities of learning. This advantage enables the deep CNNs to enhance conventional photometric methods with one single image rather than multiple ones. Saxena *et al.* [24] used discriminatively-trained Markov Random Field model to estimate depth from a single monocular image. Ladicky *et al.* [25] generalized the depth estimation and semantic segmentation as a multiple semantic classification problem. Eigen *et al.* [11] employed two (coarse and fine) deep network stacks to generate a coarse global estimation firstly and refined this estimation locally afterward. Liu *et al.* [12] formulated depth estimations into a continuous conditional random field learning problem, and presented a deep convolutional neural field model to solve the Maximum A Posteriori problem for depth estimations. Yoon *et al.* [26] adopted a Generative Adversarial Network (GAN) for fine-scale normal estimation using a single near-infrared (NIR) image. Tatarchenko *et al.* [27] used an encoder-decoder network for predicting the depth map from an RGB image.

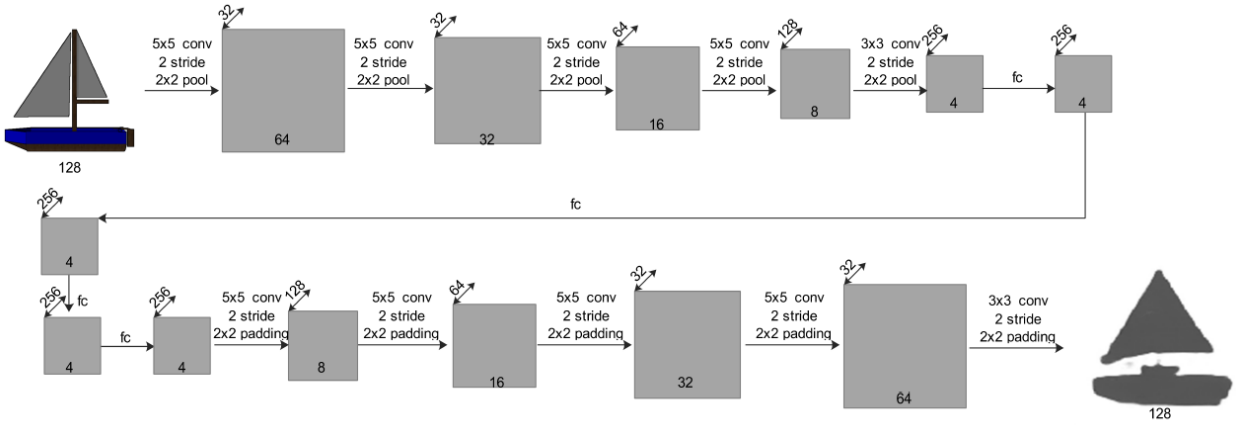

Figure 1. Model architecture.

Though the deep CNNs have high learning capability, such as estimating depth from a single image is still an ill-posed problem, the depth estimations by deep CNNs are still not accurate enough for some applications. Furthermore, the great demand for the training data makes deep CNN less practical than traditional photometric methods. Different from the existing works, we combine the deep CNN with multi-spectral PS for depth estimation. Therefore, our method can estimate the depth information from a single image with a higher precision.

In this section, the details of the proposed depth estimation scheme will be discussed. The scheme consists of two main parts: (a) a deep convolutional neural network and (b) a multi-spectral PS algorithm. The proposed method uses the multi-spectral photometric stereo to enhance the depth estimation from the deep convolutional neural network to reconstruct fine details.

3.1. Deep Convolutional Neural Network

The network is established to estimate a global depth map from a single image.

3.1.1. Architecture. The architecture of the proposed network is shown in Fig.1. The network contains fourteen layers with different weights. First, five convolutional layers are utilized for extracting depth features of an object from the input image. Then, four layers are following as fully-connected ones. The last five layers are deconvolutional ones for parsing these abstracted depth features and generating the final depth map. Except for the last deconvolutional layer, the Leaky ReLU non-linearity with the negative slope 0.2, known as the $tanh$, is applied to all the output of every layers.

The first four convolutional layers respectively filter the input with 32, 32, 64 and 128 kernels of the same 3×3 window size using the same stride of 2 pixels. The fifth convolutional layer uses 256 kernels of size 5×5 to filter the response-normalized output from the former convolutional layers. Each of the fully-connected layers have 1024 neurons. Compared with the convolutional layers, the deconvolutional layers operate with upsampling operation rather than pooling. We use a 2×2 patch that contains the target pixel in the top-left corner and pads zero elsewhere. We set the number of kernels of these five deconvolutional layers in a reverse order of the one from the convolutional layers. And the filter size of each deconvolutional layers is 3×3 as well.

3.1.2. Training. The lack of data makes difficult to train the network with real object. Therefore, our network is trained with synthetic data from the ShapeNet dataset [28]. The dataset contains 55 common object categories with about 51,300 unique 3D models. We followed the procedure described in the paper [27] for rendering 3D models to training images on the fly.

Since the trained model needs to be applied to real world data and combine the output with multi-spectral PS algorithm, we fix the viewpoint at side view and randomly sample a amount of data with light sources from 2 to 4 at random locations and with random intensity as described in [29].

Assuming a set of single image $x_i \in \{x_1, x_2, \cdots, x_n\}$ from training set with their corresponding ground truth depth maps $d_i \in \{d_1, d_2, \cdots, d_n\}$, the loss function L for the training process is as following:

$$L = \sum_{i}^{n} \|d_i - \hat{d}_i\|_1, \tag{1}$$

where $\hat{d}_i$ is the output of the network. The l_1 norm is used to represent the difference between the result of our network and the ground truth.

3.2. Multi-Spectral Photometric Stereo

To obtain the fine and detailed depth information for real objects, the propose method requires a orthographic camera

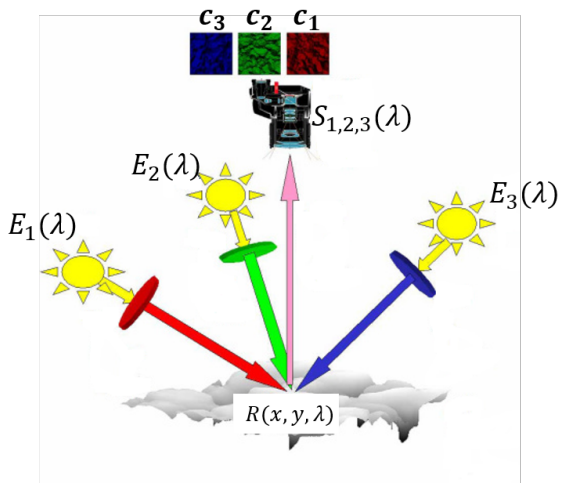

Figure 2. Configuration. A schematic representation of our multi-spectral PS technique.

and three colored light sources. As shown in Fig.2, with the projections of red, green and blue light rays from three different directions, a Lambertian surface reflects each of those colored light rays simultaneously without mixing the frequencies of light spectrum. These reflection can be sensed and separated by the RGB three-channel camera sensor such as Bayer filter sensor, Foveon X3 sensor and 3CCD.

The pixel intensity $c_i(x, y)$ of pixel (x, y) for the i-th channel is given by

$$c_i(x,y) = \sum_j (\boldsymbol{l}_j^\top \boldsymbol{n}) \int E_j(\lambda) R(x,y,\lambda) S_i(\lambda) \mathrm{d}\lambda. \quad (2)$$

The vector $\boldsymbol{n}$ is a local normal for the pixel (x, y), and $\boldsymbol{l}_j$ defines the directions of the three light rays. S_i and E_j denote the sensor sensitivity and spectral distribution for per channel i and light source j. Let the reflectance function $R(x, y, \lambda) = \rho(x, y)\alpha(\lambda)$, and $v_{ij} = \int E_j(\lambda)\alpha(\lambda)S_i(\lambda)\mathrm{d}\lambda$, where $\rho(x, y)$ is the albedo of the surface at point (x, y), and $\alpha(\lambda)$ is the characteristic chromaticity of the surface. By vectorizing v_{ij} to $\boldsymbol{v}_j = (v_{1j}\ v_{2j}\ v_{3j})$, the vector of the three channel responses at a pixel is given by

$$\boldsymbol{c} = \rho[\boldsymbol{v}_1\ \boldsymbol{v}_2\ \boldsymbol{v}_3][\boldsymbol{l}_1\ \boldsymbol{l}_2\ \boldsymbol{l}_3]^\top \boldsymbol{n}. \quad (3)$$

Let $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \boldsymbol{v}_3]$ and $\boldsymbol{L} = [\boldsymbol{l}_1\ \boldsymbol{l}_2\ \boldsymbol{l}_3]^\top$, Eq.3 can be written as:

$$\boldsymbol{c} = \rho \boldsymbol{V} \boldsymbol{L} \boldsymbol{n}. \quad (4)$$

If matrix $\boldsymbol{M} = \rho \boldsymbol{V} \boldsymbol{L}$ is known, we can compute the normal as

$$\boldsymbol{n} = \boldsymbol{M}^{-1} \boldsymbol{c}.$$

Then, the depth information can be integrated from the normal according to the definition of normal

$$\boldsymbol{n} = \frac{1}{\sqrt{1 + |\nabla z|^2}} \begin{pmatrix} \nabla z \\ -1 \end{pmatrix}, \quad (5)$$

where $z = z(x, y)$ denotes a height function of surface in front of the camera, and ∇z is the gradient of the function with respect to x and y.

3.3. Combination of DCNN and PS

Generally, the matrix $\boldsymbol{M}$ is calibrated by measuring the RGB responses corresponding to each direction of the surface. However, this calibration process requires an extra mirror sphere for estimating the light source directions with three more sequences of images. In this paper, we abandon this complex calibration process and use the output of the deep convolutional neural network discussed in the previous section. According to Eq.4, $\boldsymbol{M}$ can be expressed as

$$\boldsymbol{M} = [\boldsymbol{c}_1\ \boldsymbol{c}_2\ \boldsymbol{c}_3][\boldsymbol{n}_1\ \boldsymbol{n}_2\ \boldsymbol{n}_3]^{-1}$$

The local normal and intensity are known for 3 pixels that have equal albedo. Therefore, the problem will be solved if such 3 pixels and their normal can be found.

A normal $\boldsymbol{n}_d$ can be calculated from the depth image generated by the deep CNN. Despite the fact that the geometry structure acquired using deep CNN is not precise enough, there are still some effective depth pixels correctly estimated. We use random sample consensus (RANSAC) algorithm for choosing those effective pixels and estimating the matrix $\boldsymbol{M}$. To fulfill the aforementioned assumption, the image is segmented into different superpixels using simple linear iterative clustering (SLIC) technique, and every pixel in the same superpixel is assumed to have equal albedo and chromaticity.

With the estimated matrix $\boldsymbol{M}$, a refined and detailed depth map can be obtained from a single RGB image with the uncalibrated light sources.

4. Experiments

In the experiments of this paper, a IDS UI-358xCP-C camera is utilized for imaging. It is located in the center of a circular framework with three light sources of red, green and blue around it. The objects are placed at approximate distance of 60 cm in front of he camera. The size of the target objects are about 10 cm to ensure their visibility in the FOV of the camera. The experiments consist of two part: (a) the intermediate results of the DCNN based estimation and (b) the final results of the proposed method.

4.1. Results of Deep CNN

We use Tensorflow for implementing and training the proposed network with a graphic card of Nvidia GT730. The training process uses the batches with a size of 16. The loss function is optimized using the Adam optimizer [30] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We employ a Gaussian distribution with zero mean and a standard deviation of 0.02 for initializing weights. The learning rate is 0.0001. Different from several existing works (e.g. [11]), we don't augment our training data for the network since we need to transfer our model to real object.

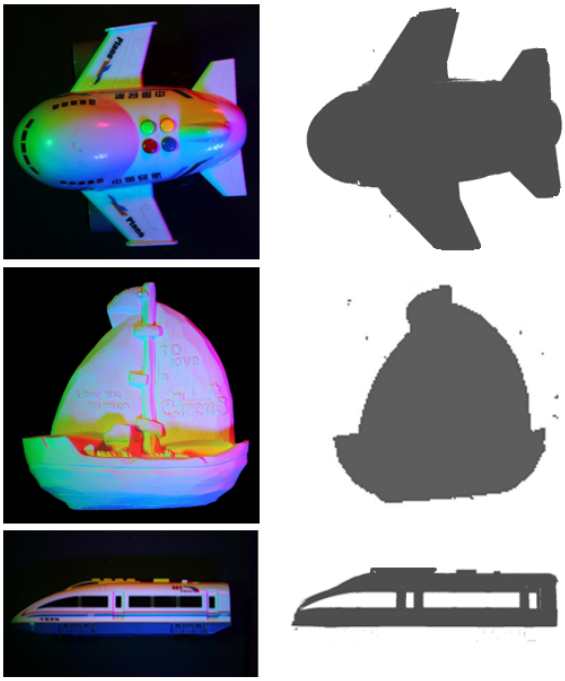

Figure 3. Results of our network. Column (a) is the images of real objects, column and column (b) is the depth estimations of our network.

We test our network with image of a plaster boat down-sampled to the size of 128×128. The estimated depth map is then combined with multi-spectral photometric stereo the the final result. The estimated depth by DCNN are shown in Fig.3.

Fig.3 demonstrates the depth estimations generated by our network. Although the depth estimation produced by our network don't include enough details of real objects, it still generates the shape and outline well.

4.2. Results of Combination

As mentioned in Section 3.3, we use the output of our DCNN network as the initial depth estimation and refine it with multi-spectral PS. Our method is tested with real objects, including a toy plane, a plaster boat and a plastic train. Each object is captured into a single image under the illuminations of three color light rays. Fig.4 shows the results of our method compared to the Kinect, deep CNN of continuous CRF [12] and multi-spectral PS. Our method has higher resolution and less cavities compared to Kinect. Combined with the depth estimation using deep CNN [12], our results can achieve finer details and obtain more accurate depth estimation. More complete outline and depth estimation can be produced by the proposed method than multi-spectral PS [20].

5. Conclusion

Estimating 3D geometry from a single image with uncalibrated illumination has always been a challenging and ill-posed problem. The deep convolutional neural networks are another solutions. However, the resolution of depth map generated by deep convolutional neural networks is less satisfied than those from conventional photometric stereo. In this paper, we enhance the depth estimation from deep convolutional neural networks using multi-spectral photometric stereo. The proposed method can well handle this ill-posed problem and enhance the result of deep convolutional neural networks at the same time.

Acknowledgments

This work is supported by International Science & Technology Cooperation Program of China (ISTCP) (No.2014DFA10410) and National Natural Science Foundation of China (NSFC) (No.41576011).

References

[1] K. Ikeuchi and B. K. Horn, "Numerical shape from shading and occluding boundaries," *Artificial intelligence*, vol. 17, no. 1-3, pp. 141–184, 1981.

[2] K. M. Lee and C.-C. Kuo, "Shape from shading with a linear triangular element surface model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 8, pp. 815–822, 1993.

[3] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical engineering*, vol. 19, no. 1, pp. 191 139–191 139, 1980.

[4] M. S. Drew and L. L. Kontsevich, *Closed-form attitude determination under spectrally varying illumination*. Simon Fraser University, Centre for Systems Science, 1994.

[5] A. Thomas, V. Ferrar, B. Leibe, T. Tuytelaars, B. Schiel, and L. Van Gool, "Towards multi-view object class detection," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1589–1596.

[6] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 519–528.

[7] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[8] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM transactions on graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 835–846.

[9] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, "Efficiently combining positions and normals for precise 3d geometry," in *ACM transactions on graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 536–543.

[10] R. Anderson, B. Stenger, and R. Cipolla, "Augmenting depth camera output using photometric stereo." in *MVA*, 2011, pp. 369–372.

[11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

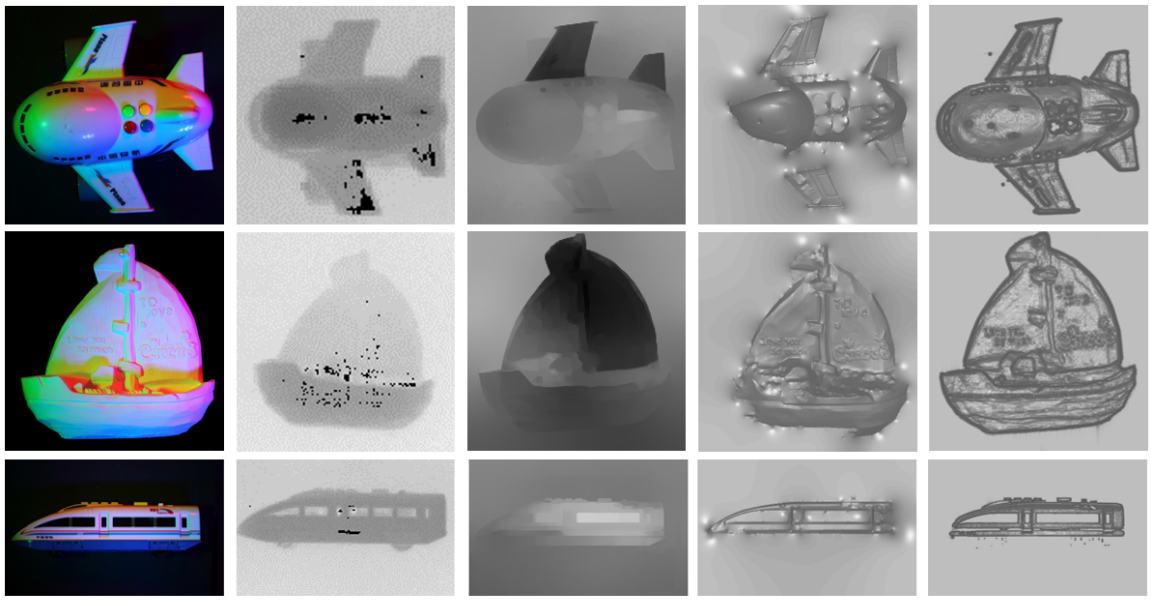

Figure 4. The final results. (a) The input images. (b) The outputs of Kinect. (c) The depth estimation of continuous CRF in deep CNN [12]. (d) The depth estimation of multi-spectral PS [20]. (e) The depth estimation of our method.

[12] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5162–5170.

[13] R. Cipolla, S. Battiato, and G. M. Farinella, *Computer Vision: Detection, recognition and reconstruction*. Springer, 2010, vol. 285.

[14] E. N. Coleman and R. Jain, "Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry," *Computer graphics and image processing*, vol. 18, no. 4, pp. 309–328, 1982.

[15] S. K. Nayar, K. Ikeuchi, and T. Kanade, "Surface reflection: physical and geometrical perspectives," DTIC Document, Tech. Rep., 1989.

[16] S. Lin and S. W. Lee, "Estimation of diffuse and specular appearance," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 855–860.

[17] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan, "A practical model for subsurface light transport," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 511–518.

[18] U. S. N. B. of Standards and F. E. Nicodemus, *Geometrical considerations and nomenclature for reflectance*. US Department of Commerce, National Bureau of Standards, 1977, vol. 160.

[19] C. Hernández, G. Vogiatzis, and R. Cipolla, "Shadows in three-source photometric stereo," in *European Conference on Computer Vision*. Springer, 2008, pp. 290–303.

[20] L. Kontsevich, A. Petrov, and I. Vergelskaya, "Reconstruction of shape from shading in color images," *JOSA A*, vol. 11, no. 3, pp. 1047–1052, 1994.

[21] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu, "Edge-preserving photometric stereo via depth fusion," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2472–2479.

[22] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin, "Shading-based shape refinement of rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1415–1422.

[23] S. Xiong, J. Zhang, J. Zheng, J. Cai, and L. Liu, "Robust surface reconstruction via dictionary learning," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 201, 2014.

[24] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *NIPS*, vol. 18, 2005, pp. 1–8.

[25] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 89–96.

[26] Y. Yoon, G. Choe, N. Kim, J.-Y. Lee, and I. S. Kweon, "Fine-scale surface normal estimation using a single nir image," in *European Conference on Computer Vision*. Springer, 2016, pp. 486–500.

[27] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Multi-view 3d models from single images with a convolutional network," in *European Conference on Computer Vision*. Springer, 2016, pp. 322–337.

[28] M. Savva, A. X. Chang, and P. Hanrahan, "Semantically-enriched 3d models for common-sense knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 24–31.

[29] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.