

Depth Estimation based on a Single Close-up Image with Volumetric Annotations

Frank P.-W. Lo, *Student Member, IEEE*, Yingnan Sun, *Student Member, IEEE*,
and Benny Lo, *Senior Member, IEEE*

Abstract—A novel depth estimation technique based on a single close-up image is proposed in this paper for better understanding of the geometric relations of a unknown scene. Previous works focus mainly on handling depth estimation problem based on global view information. Our technique, which is designed based on a deep neural network framework, utilizes monocular color images with volumetric annotations to train a two-stage neural network to estimate the depth information from close-up images. RGBVOL, a database of RGB images with volumetric annotations, has also been constructed by our group for examining the proposed methodology. Compared to previous depth estimation techniques, our method improves the accuracy of depth estimation by 4% under the condition that global cues of the scene is not available due to constrained viewing angle and distance.

I. INTRODUCTION

Depth is an important parameter to understand the geometry and structure of an unknown scene or environment. A reliable depth measurement methodology can lead to improvements in different tasks such as object grasping, target tracking, robotic navigation, image segmentation as well as 3D modeling [1], [2]. The common way of depth measurement is the use of depth sensing cameras, such as time of flight cameras or structured light 3D cameras. Despite the efficacy of depth cameras, there has been relatively few depth cameras that have been integrated into mobile or wearable devices for various reasons, such as cost and power constraints. In addition, depth cameras are often designed for limb or hand tracking with limited focal distances, and won't work with closed up views. It is for these reasons that much research on the topic of depth estimation from a monocular camera is still gaining interests. Compared with traditional depth camera techniques, estimating depth from single image is much more challenging, since it requires accurate scale calibration of the camera based on a reference object with known dimensions, such as a checkerboard, to resolve the problems of scale ambiguity, lens distortion, etc. In recent years, deep learning have been applied extensively, and much research has started to focus on using single image to predict the depth information using deep neural networks. The advantage of using deep neural network for depth prediction is that the scale and other parameters of the monocular image can be learned from the global cues of the scene without the need of camera calibration [3], [4]. For example, a set of parallel lines converges to a single

point, which called vanishing point, in the image space. This point can provide significant information on the distance of the items in the image. However, in spite of the significant advances on using various specialized architecture to handle the problem utilizing global cues such as vanishing points, line angles and perspective, these is still limited work on estimating depth information from close-up images which do not contain any global view information [5]–[7]. Global information cannot always be perfectly obtained in some situation, such as predicting depth based on a short-range image or bird-view image, which induces a significant error on depth estimation. Intuitively, to estimate the depth from a close-up image, more features are required on top of previous proposed features and architecture. In mathematics, as we all know that volume of different geometric shapes can always be rearranged to calculate depth based on their corresponding volume formulas. To leverage the coupled nature of volume and depth, we hypothesize that color RGB images with volumetric annotations can enhance the performance of depth estimation without relying much on global view information. To our best knowledge, there is no published work on using volumetric information to estimate depth in the real world scene.

This work explores a novel way of using single color images with volumetric annotations to predict the depth of the scene. To examine the feasibility of our proposed architecture, a new dataset RGBVOL¹, which is composed of 3000 RGB images with volumetric annotations, has also been constructed based on our proposed volume estimation technique [8]. The rest of the paper is organized as follows: Section II details the methodology of our volume estimation technique and the neural network architecture. Section III presents the experimental results of the proposed volume-based depth estimation methodology. Discussion and conclusion will be presented in Section IV.

II. DETAILED INFORMATION AND METHODS

A. Volume Estimation

Fig 1 shows the flow chart on the process of generating a volumemap. At the beginning, RealSense Camera SR300, a RGBD camera developed by Intel, is used to capture RGB images and construct depthmaps as shown in Fig. 2 and Fig. 3 respectively. Due to the disparity of the RGB camera and the infrared sensors on the RGBD camera, in order to construct depthmaps for close-range items, image

F. Lo, Y. Sun, and B. Lo are with the Hamlyn Center, Imperial College London, London, U.K. {po.lo15, y.sun16, benny.lo}@imperial.ac.uk

¹<https://www.doc.ic.ac.uk/~ys4315/Y/>

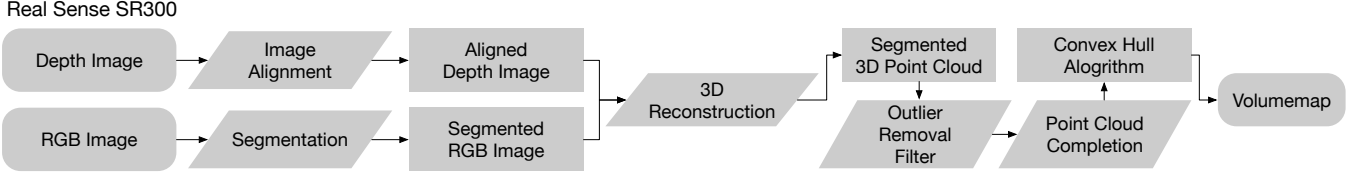


Fig. 1. Flowchart for constructing the volumemap



Fig. 2. RGB images of object replicas captured by Realsense SR300

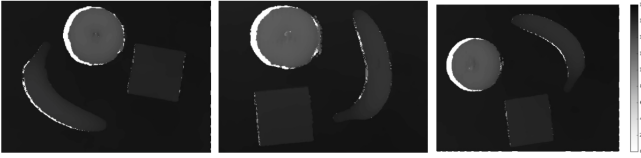


Fig. 3. Depthmap captured by Realsense SR300

alignment (also known as image registration) shall be carried out in advance to align the RGB image and the depthmap. This disparity causes significant errors in depth estimation especially if the image is captured in a short distance (< 50 cm). To align the frames, a homography, a 3×3 matrix as shown in equation 1, is required. The homography can be computed with the following steps. First, features points should be extracted with suitable descriptors. Second, the n point-correspondences (n is normally set as 4) in different frames should be determined by matching descriptors. After that, the homography matrix can then be deduced from the identified feature points. These procedures can normally be achieved using Random SAmple Consensus (RANSAC) scheme [9]–[11].

$$\begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = H \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \quad (1)$$

where (u_1, v_1) and (u_2, v_2) represent the same feature points in two different frames respectively, H refers to the homography matrix. After the depthmap is realigned with the RGB image, the next step is to perform object segmentation with the use of the aligned depth images. As this work mainly focuses on depth estimation, deep learning has not been used for segmentation. Instead, a simple Hue-Saturation-Intensity (HSV) color filters are applied for segmentation. Afterwards, segmented depthmap can be converted to heightmap, which represents the pixel-wise height of the image, with the use of the following equations.

$$heightmap(u, v) = depthmap(u, v) - depthmap_g(u, v) \quad (2)$$

where $depthmap_g(u, v)$ refers to depth of the surface the object is placed. Since the image capturing angle cannot always be perfectly perpendicular to the surface, the depth of the surface varies pixel by pixel. More importantly, due to the limited view from a single RGB image, a significant portion of the object cannot be seen from the image and depth cannot be obtained. It is for this reason that matrix completion technique is applied to estimate the unavailable depth information. Due to the high correlation between pixels, k-nearest-neighbors (kNN) has been applied to replace the missing depth pixel in order to construct the $depthmap_g$. After the heightmap is generated using equation 2, it is then converted into world coordinate (also known as 3D point cloud). In 3D reconstruction, depthmaps and the focal length of the depth camera are required for the reconstruction. Focal length can be found in the intrinsic matrix after camera calibration or estimated based on field of view (FOV) of the camera. The world coordinated can then be converted as follows:

$$\begin{aligned} Z &= heightmap(u, v) + \sum_u^h \sum_v^w depthmap_g(u, v) / (w * h) \\ X &= (u - Center_u) * Z / focalLength \\ Y &= (v - Center_v) * Z / focalLength \end{aligned} \quad (3)$$

where u, v refer to the coordinates in the image and X, Y and Z refer to the coordinates in the world coordinate. $Center_u$ and $Center_v$ refer to the coordinates of the image center (320 pixel and 240 pixel are chosen in the case of using 640×480 as image dimension). $focalLength$ represents the focal length of the lens when the object is in focus (focal length of Realsense Depth Camera SR300 is found to be 618 pixel after calibration). The projected points in the world coordinates are shown in Fig.4.

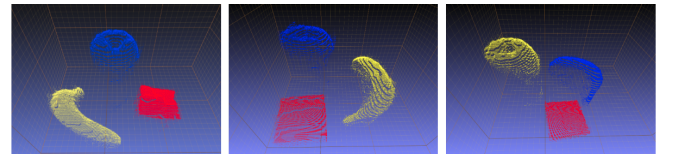


Fig. 4. 3D point cloud generated by the heightmap

The point cloud generated by a single view (top view) depth image can only be used to handle objects with narrow top and wide bottom since infrared light generated by depth camera cannot reach the bottom if the upper part is wider. Due to the reason of limited viewing angle, this will induce error on volume estimation. To tackle this problem, several

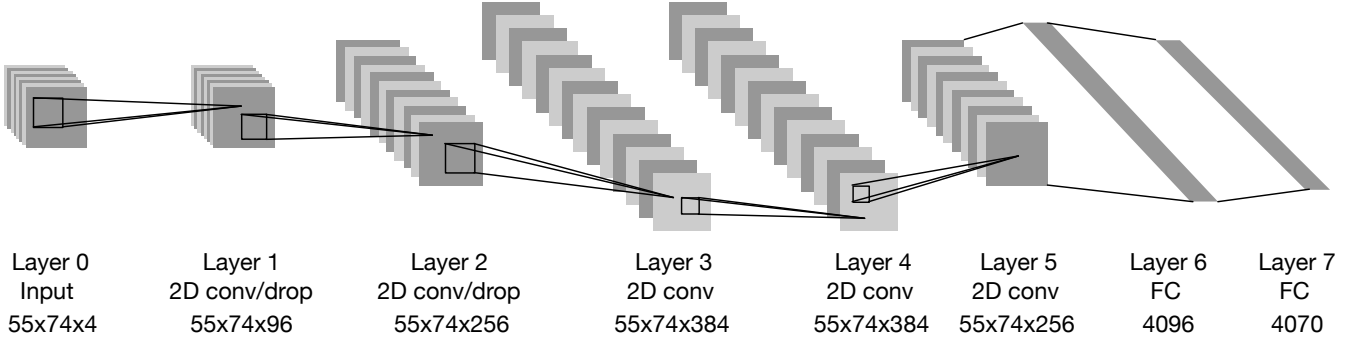


Fig. 5. The neural architecture proposed for single image depth estimation

ways are proposed by previous works to use multiple images from different angles to reconstruct 3D point cloud or carry out point cloud completion based on symmetry [8]. Once the 3D point cloud has been preprocessed and determined, the volume can then be computed based on the convex hull algorithm, a 3D mesh reconstruction technique. Convex hull C of a point set P in the Euclidean space is the smallest convex sets which is able to enclose all the points in P . The detailed information and methodology regarding point cloud completion and convex hull algorithm have been discussed in our previous research paper [8]. For the feasibility of constructing a database, including images all labeled with their volume, for further deep learning training processes, single view (top view) has been applied solely in this paper, without using multiple viewing angles, to estimate the volume of the item. Once the volume has been estimated, the volume will be labeled on corresponding items pixel by pixel to construct the volumemaps as shown in Fig.6. The dataset RGBVOL can then be constructed for the training process.



Fig. 6. Volumemap generated by convex hull algorithm

B. Neural Network Architecture

Similar to the modified CNN-based neural network by Eigen et al. [7]. for single image depth prediction, the topology of our neural network, shown in Fig. 5, has been designed empirically based on our new dataset of images with volumetric annotations. Our proposed methodology is composed of two neural networks, also can be shown as a two stage process, including the network for volumemap estimation based on input RGB images (first stage) and the network for further depth estimation (second stage). The network is composed of 5 convolutional layers, several dropout layers and 2 fully connected layers which are combined in one to form a unified neural network.

For the first stage, it can be known as a modified segmentation process. The RGB images are passed into the first

convolutional layer $L1$. As shown in Fig. 5, the filters used in the first convolutional layer is in the form of $15 \times 15 \times 3$ (3 here refers to the 3 RGB channels) and the stride is set to 2×2 . With the use of same padding, the feature map has the same dimension as the input. Referring to network proposed by Eigen et al. [7], 96 filters have been applied for the first layer so that the output will be of size of $55 \times 74 \times 96$. The feature maps will then be directed into the Rectified Linear Unit (ReLU) layer, 2×2 max-pooling layer and dropout layer to prevent overfitting. For $L2$, the $55 \times 74 \times 96$ matrix has been passed into another convolutional layer with filter size in $5 \times 5 \times 96$, and the stride is set to 2×2 again. Since the number of filter chosen for the second layer is 256, the output is shown as $55 \times 74 \times 256$. For $L3$, $L4$ and $L5$, they are all designed as convolutional layers with $3 \times 3 \times 384$, $3 \times 3 \times 384$, $3 \times 3 \times 256$ as their filter sizes followed by activation function, max pooling layers and dropout layers as well. After passing through 5 convolutional layers, the feature maps have been fed into 2 fully connected layers with 4096 and 4070 neurons respectively. The output of the deep learning network is a volumemap of size 55×74 . In the second stage, volumemaps, with volume corresponding to each item labeled, generated by the previous stage will be fused with the RGB images to estimate the depth which means the input size of the neural network will be in the form of $55 \times 74 \times 4$. The neural network used for training for this stage is similar to the previous one but with different loss function. The mathematical expression for the loss function is shown as follows:

- Loss function for volumemap estimation: L2 norm

$$\sum_u^h \sum_v^w (v(\tilde{u}, v) - v(u, v))^2 \quad (4)$$

where $v(\tilde{u}, v)$ and $v(u, v)$ refer to the pixel from the estimated volumemap and the ground truth of volumemap respectively. w and h refer to the width and height of the image respectively.

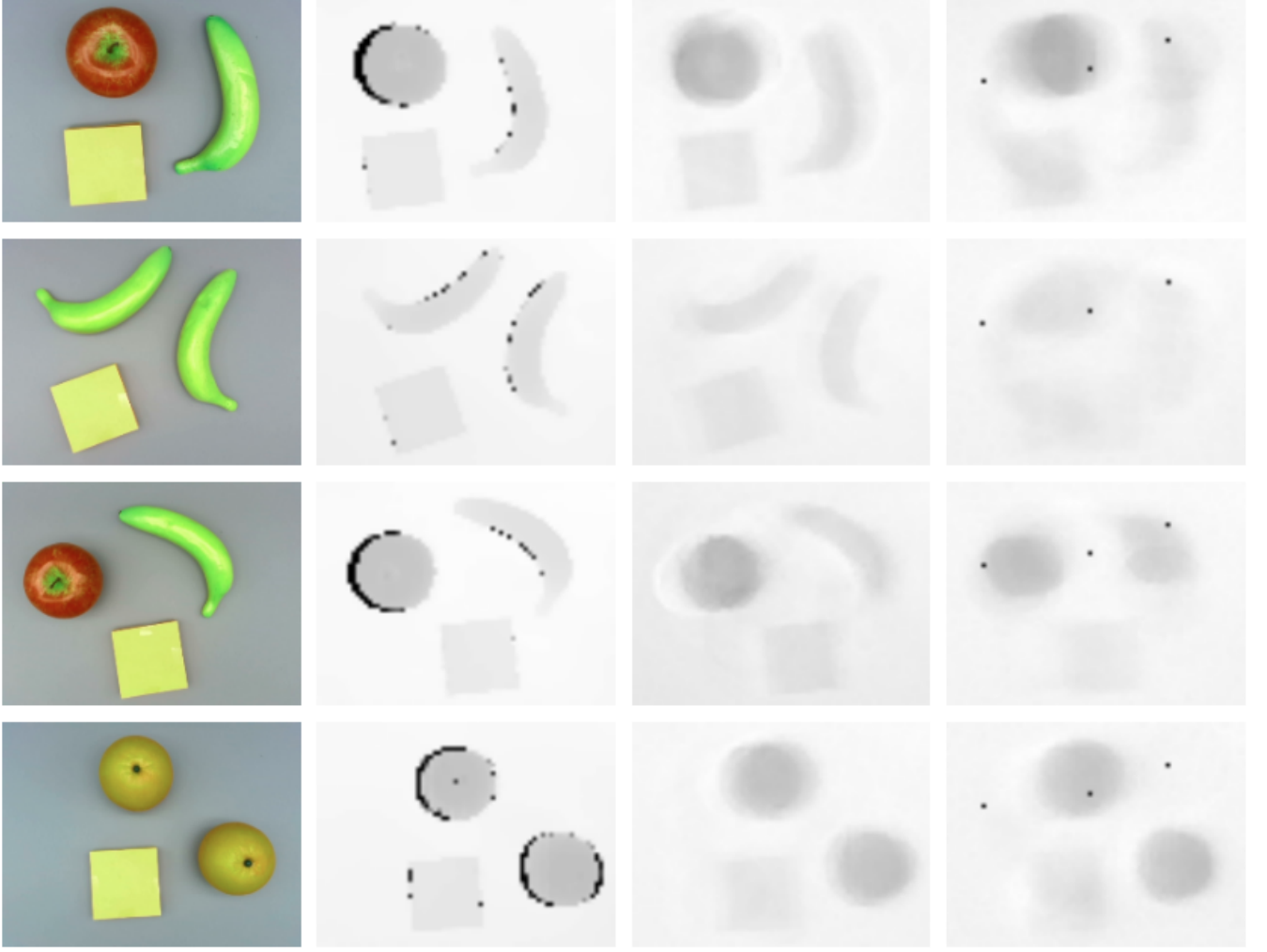


Fig. 7. Experimental results generated by the two-stage neural network (a) RGB image (b) Ground Truth (c) Our proposed technique (d) Eigen et al. method

- Loss function for depth estimation: scale-invariant error [7]

$$\frac{1}{w * h} \sum_u \sum_v^h (d(\tilde{u}, v) - d(u, v))^2 - \frac{1}{2(w * h)^2} \left(\sum_u \sum_v^h (d(\tilde{u}, v) - d(u, v)) \right)^2 \quad (5)$$

where $d(\tilde{u}, v)$ and $d(u, v)$ refer to the pixel from the estimated depthmap and the ground truth of depthmap respectively. w and h refer to the width and height of the image respectively.

III. EXPERIMENTAL RESULTS

To examine the feasibility of our proposed architecture, a new dataset, RGBVOL¹, has been constructed for training. The dataset is composed of 14 plastic artificial fruits and other objects with different volume recorded under varying

camera height and arrangement. 3000 images have been recorded in total and labeled with their corresponding volume.

1) *Accuracy of volume estimation based on CNN:* Comparing the volume estimation result of our approach with the ground truth, the accuracy of volume estimation method is shown in Table I. The accuracy of the volume varies for different objects. It achieves an accuracy up to 97.87%, 97.75%, 98.21%, 99.49%, and 99.36% for apple, orange, banana, memo pad and mobile phone respectively.

TABLE I
ERROR IN VOLUME ESTIMATION FOR DIFFERENT ITEMS

	Error in Volume Estimation				
	Apple	Orange	Banana	Memo	Mobile Phone
Mean	2.13%	2.25%	1.79%	0.51%	0.64%
S.D.	0.31%	0.24%	0.21%	0.12%	0.15%

2) *Depth Estimation based on Volumetric Annotations:* Three different accuracies, including depth accuracy of the item, depth accuracy of the background and the depth

¹<https://www.doc.ic.ac.uk/~ys4315/Y/>

TABLE II
ACCURACY IN DEPTH ESTIMATION BY USING DIFFERENT KERNEL SIZES

Kernel size (first layer)	Accuracy in depth estimation							
	5x5x3	5x5x4	10x10x3	10x10x4	15x15x3	15x15x4	25x25x3	25x25x4
Method	RGB	RGBV	RGB	RGBV	RGB	RGBV	RGB	RGBV
Depth accuracy_total (train)	0.958	0.966	0.958	0.967	0.955	0.964	0.952	0.964
Depth accuracy_item (train)	0.911	0.931	0.911	0.931	0.905	0.929	0.898	0.927
Depth accuracy_background (train)	0.962	0.973	0.962	0.971	0.959	0.971	0.961	0.972
Depth accuracy_total (test)	0.919	0.931	0.918	0.932	0.913	0.931	0.906	0.934
Depth accuracy_item (test)	0.847	0.877	0.845	0.876	0.837	0.877	0.832	0.876
Depth accuracy_background (test)	0.927	0.940	0.926	0.942	0.921	0.942	0.909	0.943

TABLE III
ACCURACY IN DEPTH ESTIMATION BY TRIVIAL SHIFT OF THE VOLUME LABELS

	Accuracy in depth estimation							
	<1%	1.25%	2.5%	5%	8%	15%	30%	50%
Depth accuracy_total (test)	0.929	0.925	0.913	0.870	0.821	0.727	0.534	0.434
Depth accuracy_item (test)	0.876	0.873	0.876	0.869	0.817	0.732	0.541	0.442
Depth accuracy_background (test)	0.937	0.932	0.916	0.871	0.822	0.726	0.531	0.431

accuracy of the whole image, have been used to evaluate the performance of our proposed method and the following equations are used to calculate the accuracy:

- Depth accuracy of the item: the accuracy of depth estimation with volumetric annotations

$$\frac{\sum_u^h \sum_v^w (d(\tilde{u}, v) - d(u, v)) * mask(u, v)}{\sum_u^h \sum_v^w (d(u, v) * mask(u, v))} \quad (6)$$

where $d(\tilde{u}, v)$ and $d(u, v)$ refer to the estimated depthmap and the ground truth of the depthmap respectively. $mask(u, v)$ refers to the binary mask of the image with background labeled as 0.

- Depth accuracy of the background: the accuracy of depth estimation without volumetric annotations

$$\frac{\sum_u^h \sum_v^w (d(\tilde{u}, v) - d(u, v)) * (1 - mask(u, v))}{\sum_u^h \sum_v^w (d(u, v) * (1 - mask(u, v)))} \quad (7)$$

- Depth accuracy of the whole image: the accuracy of depth estimation of the whole image

$$\frac{\sum_u^h \sum_v^w (d(\tilde{u}, v) - d(u, v))}{\sum_u^h \sum_v^w d(u, v)} \quad (8)$$

The results of the proposed volume-based depth estimation are compared with the state-of-the-art single image depth estimation technique based on the dataset RGBVOL, as shown in Fig. 7. The performance of our proposed architecture, when using different kernel size in filters, has been listed in Table II. As shown in the results, the newly proposed technique outperforms previous proposed methods. The most

significant improvement lies on the depth estimation accuracy of the object which increases by 3.0%, 3.1%, 4.0% and 4.4% when kernel sizes increase. From the results, the proposed methodology benefits the most from increasing the kernel sizes and not only that, the testing accuracy of the background also increases by 1.3%, 1.6%, 2.1% and 3.4% respectively. From this increasing trend, it proves that the objects with volumetric annotations can provide scale and distance information of the scene. The experimental results support our hypothesis that geometric shapes and volume of objects can be used to improve depth estimation. Moreover, experiments have been carried out to examine the robustness of the method and test whether the performance of depth estimation will be affected by trivial shift of the volume labels in the first stage of the processing. Random noise have been added into volumemaps to induce errors. The depth estimation drops by an average of 0.26%, 0.6%, 1.8% and 4.4% when the noise error fed into the volumemap were 0.5%, 1.25%, 2.5% and 5% respectively. The detailed experimental results are shown in Table III.

IV. DISCUSSION AND CONCLUSION

Depth estimation from a single close-up image is a challenging research work due to the reason of lacking global cues of the environment. In this work, we demonstrated that the volumetric information provides the network the scale of the image and objects, and improves depth estimation as shown in the experimental results. In future work, we are planing to extend the method to incorporate further information which can be obtained from a single image for better depth estimation, such as the semantic labels and geometric shapes. Furthermore, we are also planned to incorporate our proposed technique on top of the neural architecture with conditional random field to investigate if it will further improve the volume and depth estimation.

ACKNOWLEDGMENT

This project is partially supported by Lee Family Scholarship, awarded to Frank P.-W. Lo and the Innovative Passive Dietary Monitoring Project funded by the Bill & Melinda Gates Foundation (Opportunity ID: OPP1171395).

REFERENCES

- [1] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Computer Vision - ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 746–760, ISBN: 978-3-642-33715-4.
- [2] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2009, ISSN: 0162-8828 VO - 31. DOI: 10.1109/TPAMI.2008.132.
- [3] R. Memisevic and C. Conrad, "Stereopsis via deep learning," in *NIPS Workshop on Deep Learning*, 2011.
- [4] K. R. Konda and R. Memisevic, "Unsupervised learning of depth and motion," *CoRR*, vol. abs/1312.3, 2013. arXiv: 1312 . 3429. [Online]. Available: <http://arxiv.org/abs/1312.3429>.
- [5] X. Ma, Z. Geng, and Z. Bie, "Depth Estimation from Single Image Using CNN-Residual Network," in *SemanticScholar*, 2017.
- [6] E. J. Chappero, R. A. Guerrero, and F. J. Sern, "Vanishing Point estimation from monocular images," in *SemanticScholar*, 2010.
- [7] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," *CoRR*, vol. abs/1406.2, 2014. arXiv: 1406.2283. [Online]. Available: <http://arxiv.org/abs/1406.2283>.
- [8] A. Gao, F. P. W. Lo, and B. Lo, "Food volume estimation for quantifying dietary intake with a wearable camera," in *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2018, pp. 110–113, ISBN: VO -. DOI: 10.1109/BSN.2018.8329671.
- [9] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, ISSN: 0001-0782. DOI: 10.1145/358669.358692. [Online]. Available: <http://doi.acm.org/10.1145/358669.358692>.
- [10] O. Chum, "Two-view geometry estimation by random sample and consensus," *PhD Dissertation, CTU*, 2005.
- [11] S. Choi, T. Kim, and W. Yu, "Performance Evaluation of RANSAC Family," in *British Machine Vision Conference (BMVC)*, 2009.