2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

360SD-Net: 360° Stereo Depth Estimation with Learnable Cost Volume

Ning-Hsu Wang[1], Bolivar Solarte[1], Yi-Hsuan Tsai[3]
Wei-Chen Chiu[2], Min Sun[1]
[1]National Tsing Hua University, [2]National Chiao Tung University, [3]NEC Labs America
albert100121@gapp.nthu.edu.tw, enrique.solarte.pardo@gmail.com, wasidennis@gmail.com,
walon@cs.nctu.edu.tw, sunmin@ee.nthu.edu.tw

***Abstract*— Recently, end-to-end trainable deep neural networks have significantly improved stereo depth estimation for perspective images. However, 360° images captured under equirectangular projection cannot benefit from directly adopting existing methods due to distortion introduced (i.e., lines in 3D are not projected onto lines in 2D). To tackle this issue, we present a novel architecture specifically designed for spherical disparity using the setting of top-bottom 360° camera pairs. Moreover, we propose to mitigate the distortion issue by (1) an additional input branch capturing the position and relation of each pixel in the spherical coordinate, and (2) a cost volume built upon a learnable shifting filter. Due to the lack of 360° stereo data, we collect two 360° stereo datasets from Matterport3D and Stanford3D for training and evaluation. Extensive experiments and ablation study are provided to validate our method against existing algorithms. Finally, we show promising results on real-world environments capturing images with two consumer-level cameras. Our project page is at https://albert100121.github.io/360SD-Net-Project-Page.**

I. Introduction

Stereo depth estimation is a long-lasting yet important task in computer vision due to numerous applications such as autonomous driving, 3D scene understanding, etc. Despite the majority of studies are for perspective images, disparity can be defined upon various forms of image pairs. For instance, the human binocular disparity is defined as the angle difference between the point of projection on the retina, which is part of a sphere rather than a plane. Similar to human vision, the angle difference of a pair of 360° cameras with spherical projection can also be defined as disparity (see Fig. 1(a)). By taking the advantage of 360° cameras for having a complete observation in an environment, the stereo depth estimation obtained from these cameras enables the 3D reconstruction of the entire surrounding. This is a powerful advantage for advanced applications, e.g., 3D scene understanding.

In this paper, we aim to estimate stereo depth information from a pair of equirectangular images (see Fig. 1(b)(c)), in which they are used in most consumer-level 360° cameras. For simplicity, we thereafter refer equirectangular images to as 360° images. The critical issue needed to cope with is the severe distortion introduced in the process of equirectangular projection. First, horizontal lines in 3D are not always the lines in 2D when we use 360° cameras. This implies that the typical configuration of the left-right stereo rig may not preserve the same property of epipolar lines. Therefore, we

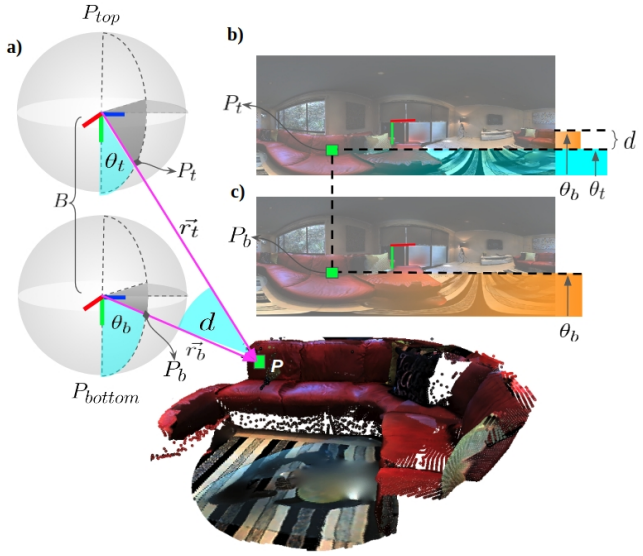

Fig. 1: Spherical disparity under (a) top-bottom camera pairs (P_{top} and P_{bottom}) with baseline B. Panel (b)(c) show top and bottom equirectangular projections, respectively. P_t and P_b are projection points from a 3D point onto the spherical surface (a) and equirectangular coordinate (b)(c). $\vec{r}_t$ and $\vec{r}_b$ are projection vectors for the top and bottom cameras, respectively. θ_t and θ_b are the angles between the south pole and its respective projection vector. $d = \theta_b - \theta_t$ is the angular disparity. In panel (b)(c), the 3D point projects to the same horizontal position but different vertical positions reflecting the disparity.

configure two cameras in a top-bottom manner, such that the epipolar lines on a pair of images are vertically aligned (see Fig. 1). Second, pixels near the top and bottom of the images are stretched more than those located around the equator line. Hence, the corresponding patches at different vertical locations are likely to have different visual characteristics due to different levels of distortion. This encourages us to propose a novel framework for learning correspondence in top-bottom aligned equirectangular images.

We demonstrate the benefit of each component through extensive ablation study and compare the performance with deep-learning baselines (i.e., PSMNet [1] and GCNet [2]) and conventional stereo matching approaches (i.e., ASW [3], Binocular [4], Kim's [5]). The efficacy of our full model is validated in improving depth estimation for 360° stereo

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

cameras on two synthetic datasets, as well as generalization to real-world images. The main contributions are as follows:

- Propose the first end-to-end trainable network for stereo depth estimation using 360° images.
- Develop a series of improvements over existing methods to handle the distortion issue, including the usage of polar angle.
- Propose a novel learnable shifting filter for building the cost volume which is empirically better than standard pixel-shifting in the spherical projection.
- Introduce our 360° stereo dataset collected from Matterport3D [6] and Stanford3D [7], composed of equirectangular pairs and depth/disparity ground truths.
- Generalize to real-world environments using two consumer-level 360° cameras with a model trained on the synthetic dataset.

II. Related Work

A. Classical Methods

Prior to the recent advances of deep learning, numerous research efforts have been devoted to stereo matching and depth estimation. These classical stereo matching algorithms can be roughly categorized into local and global methods. In general, global methods (e.g., Semi-Global Matching (SGM) [8]) are able to estimate a better disparity map, but they need to solve a complicated optimization problem. On the other hand, local algorithms (e.g., Adaptive Support-Weight approach (ASW) [3], [9] and Weighted Guided Image Filtering (WGIF) [10]) are faster and widely used in many embedded applications, but they suffer from the aperture problem or ambiguous matches on homogeneous regions.

Regarding 360-view methods, Kang *et al.* [11] target at stereo 360° images on a cylinder projection, but do not consider a full 360-view (4π steradians). In addition, Im *et al.* [12] tackle monocular 360° depth estimation using structure-from-motion and sphere sweeping algorithm. Although this model leverages a spherical projection, it is limited to handle short sequences with a high computational cost. Similar to our setting, Li [4] presents a top-bottom camera setting to define spherical disparity, while Kim *et al.* [5] follow the same camera setting but with a PDE-based regularization method to refine the disparity results. Although these methods tackle 360° stereo depth estimation directly on spherical projection, they still encounter problems of ambiguous matches, artifacts, or diffused surfaces, where we address them via designing a learning-based framework.

B. Deep Learning-based Stereo Method

Recently, deep learning techniques achieve great progress on stereo depth estimation. These techniques can be summarized as a framework with four main components: (1) feature extraction, (2) cost aggregation, (3) cost volume construction, and (4) disparity optimization. For instance, Koch [13] and Zbontar *et al.* [14] use a deep metric learning network (e.g., Siamese network) to focus on learning a feature representation in order to obtain better matching cost. Furthermore, Luo *et al.* [15] speed up the computation by replacing the concatenation with inner-product for cost aggregation on deep features extracted from stereo pairs. Considering full-trainable models, GCNet [2] proposes an end-to-end deep network, which has a multi-scale 3D convolution module for producing a more robust disparity regression. Moreover, PSMNet [1] steps further to have spatial pyramid pooling for taking global context information into cost volume and equipping the 3D convolution with a stack of hourglass network to achieve better disparity estimation. Despite the high performance of the mentioned approaches on stereo perspective views, they do not output desirable results using 360° images since properties such as distortion are not considered in their model design.

C. Vision Techniques for 360°Camera

When the consumer-level 360° cameras were made easily available and affordable, it attracts significant research interest from the computer vision and robotics communities. For instance, Cohen *et al.* [16] and Esteves *et al.* [17] process spherical information on spectral-domain for classification, whereas KTN [18] and Flat2sphere [19] focus on designing spherical convolution kernels such that the network can support multiple recognition tasks in 360° images. On the other hand, several works [20], [21], [22] leverage 360-views to reconstruct layout scenes from equirectangular images as input. Similarly, [23], [24] address the problem of saliency detection in 360° videos for exploring the rich content of a scene in a more efficient manner using the full view of equirectangular representation and dealing with distortion properly. For depth estimation purposes, [25], [26], [27] tackle monocular depth estimation from 360° images via leveraging re-projection models, rendered scenes, and structures-from-motion techniques. Recently, SweepNet [28] targets multi-view stereo depth estimation applying a deep network on four fish-eye images re-projected into concentric virtual spheres to estimate 360° depths.

Despite the previous approaches, there exists a literature gap in 360° stereo-depth estimation using convolutional networks. Therefore, we provide a novel deep network, which relies on two equirectangular images as input and deals with distortion effectively. Such input is the minimum requirement for a stereo setup that keeps the benefits of a full 360° view [5], [4]. Moreover, our proposed model is capable of being applied directly by commercial-level 360° cameras, making this solution highly affordable. To the best of our knowledge, we are the first to target at deep learning-based stereo depth estimation from 360° images. We note [29], [30] as concurrent works with similar idea to our paper.

III. Method

The proposed framework, namely 360SD-Net, investigates a unique stereo depth estimation pipeline for 360° cameras. We first introduce our camera setting and define the spherical disparity. Then, we propose the end-to-end trainable model as depicted in Fig. 2.

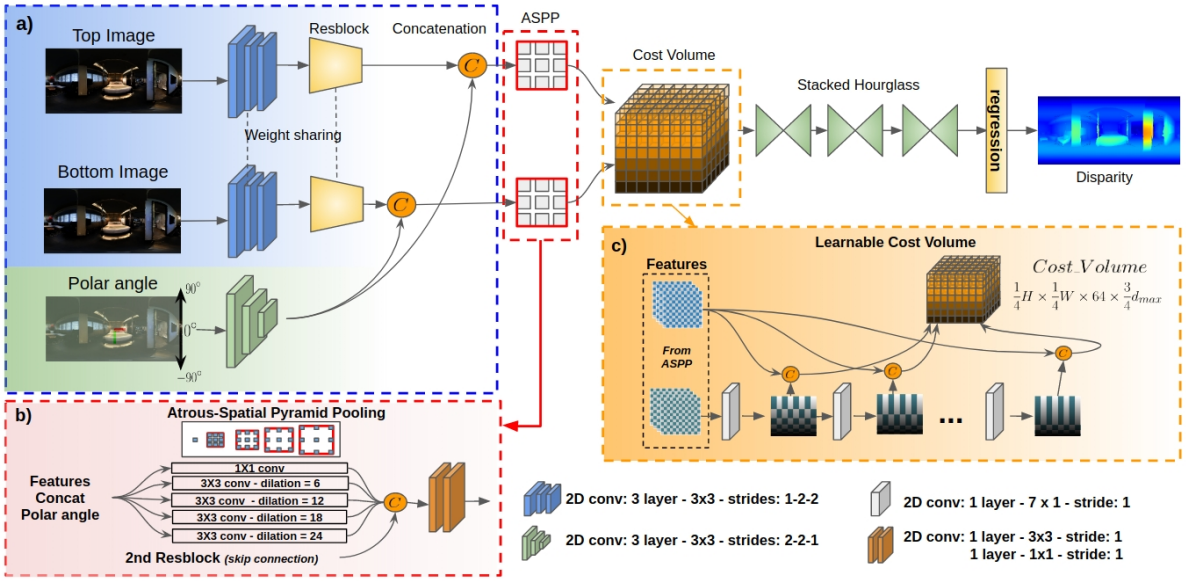

Fig. 2: Our network mainly consists of three parts: a) two-branch feature extractor that concatenates the stereo equirectangular images and the polar angle in a late fusion setting, b) the ASPP module to enlarge the receptive field, and c) the learnable cost volume to account for the nonlinear spherical projection. Finally, we use the Stacked-Hourglass module to output the final disparity map.

A. Camera Setting and Spherical Disparity

We use a top-bottom camera setting (similar to [4], [5]), where the stereo correspondence lies on the same vertical line on the camera spheres (see Fig. 1(a)). This setting also ensures that the correspondence lies on the same vertical line in 360° images captured under equirectangular projection (see Fig. 1(b,c)). Our setting can be built with relatively low cost since most consumer-level 360° cameras capture images under equirectangular projection.

We now define spherical disparity using the following terms (see Fig. 1). P_t and P_b are projection points from a 3D point P onto the camera sphere of the top and bottom camera, respectively. $\vec{r}_t$ and $\vec{r}_b$ are projection vectors, while θ_t/θ_b are the angles between the south pole and $\vec{r}_t/\vec{r}_b$ for the top and bottom cameras, respectively. The disparity is defined as the difference between the two angles with following equation $d = \theta_b - \theta_t$. The depth with respect to the top camera equals to the norm of $\vec{r}_t$, which is computed as follows,

$$|\vec{r}_t| = B \cdot \left[\frac{\sin(\theta_t)}{\tan(d)} + \cos(\theta_t) \right] , \qquad (1)$$

where B is the baseline between top and bottom cameras. Note that the disparity and depth relation is not fixed as in perspective stereo cameras, but varies according to the angle θ_t. Hence, the meaning of disparity estimation error becomes less intuitive. In practice, we mainly evaluate depth instead of disparity estimation.

B. Incorporation with Polar Angle

As described in Section II, deep stereo depth estimation methods [1], [14], [2] disregard distortion introduced in equirectangular images. To address this problem, we add the polar angle (see Fig. 2(a)) as the model input for additional geometry information since it is closely related to the distortion. In order to separate geometry information from the RGB appearance information, we apply residual blocks for RGB input and three Conv2D layers for polar angle instead of directly concatenating model input (i.e., early fusion design). Then, both outputs are concatenated after feature extraction, in which we refer to this procedure as our late fusion design. The comparison of both designs is shown in the experimental section.

C. ASPP Module

After fusing image features with the geometry information, we still have to manage the spatial relationship among pixels, since 360° images provide a larger field-of-view than regular images. In order to consider different scales spatially, we adopt recent advances ASPP [31] as proposed for semantic segmentation (see Fig. 2(b)). This module is a dilated convolution design considering multi-scale resolutions at different levels of the receptive field. In order to reduce the large memory consumption for cost volume-based stereo depth estimation, we perform random cropping during training.

D. Learnable Cost Volume

The following critical step for stereo matching is to construct a 3D cost volume by computing the matching costs at a pre-defined disparity levels with a fixed step-size. This step-size in a typical 3D cost volume is one pixel, i.e., approaches like GCNet [2] and PSMNet [1] concatenate left and right features to construct 3D cost volume based on one-pixel step-size. However, with the distortion introduced

by equirectangular projection, per-pixel step-size is not consistent with the geometry information from the polar angle input. Under this premise, we introduce a novel learnable cost volume (LCV) in our 360SD-Net using a shifting filter, which searches the optimal step-size on "degree unit" in order to precisely construct the optimal cost volume.

We design our LCV with a shifting filter via a 7×1 Conv2D layer, as shown in Fig. 2(c), and apply channel-wise shifting with the proposed filter to prevent the mixture between channels. This filter design allows vertical shifting to satisfy our stereo setting and retains the full view of the equirectangular images. Therefore, the best shifting step-size of the feature map would be learned by convolution. Note that, we apply replicated-padding instead of zero-padding before each convolution to retain the boundary information. To ensure stable training in practice, we still follow the normal cost volume shifting (freezing the parameters for the shifting Conv2D) in the first 50 epochs and start learning the cost volume shifting afterward.

E. 3D Encoder-Decoder and Regression Loss

We adopt the stacked hourglass [1] as our 3D Encoder-Decoder and the regression as in [2] to regress continuous disparity values. It is reported that this disparity regression is more robust than classification-based stereo depth estimation methods. For the loss function, we use the smooth L1 loss with the ground truth disparity.

IV. Experimental Results

A. Dataset and System Configuration

Due to the lack of 360° stereo dataset, we have collected two photo-realistic datasets MP3D and SF3D through Matterport3D [6] with Minos virtual environment [32] and re-projection of Stanford3D point clouds [7]. Considering the complexity and the extensive effort required to stitch, calibrate, and collect real-world RGB images and depth maps, which is not suitable for the training of deep models, we train our model solely on the presented synthetic data.

The setting of our dataset is a pair of 360° top-bottom aligned stereo images with equirectangular projection. The resolution of these images is 512 in height and 1024 in width, which is commonly used in 360° works [22], [24], [33]. The baseline of our stereo system is set to 20 cm, and the number of data we have collected in MP3D/SF3D datasets for training, validation, and testing are 1602/800, 431/200, 341/203, respectively. Each data consists of four components, a RGB-image pair, depth, and disparity. For data collection, we have diversified indoor scenarios in each set to prevent similarities and repetitiveness. Furthermore, the two datasets and code will be made available to the public.

We also provide qualitative results on real-world scenes to show the generalization of our model between synthetic training and real-world testing. These real-world scenes are collected with two well-known consumer-level 360° cameras, Insta360® ONE X (Fig. 3). Both cameras are calibrated using a 6x6 Aprilgrid and the toolbox calibration *Kalibr*, in particular [34]. On the other hand, to preserve our camera setting described in Section III-A, we align the polar axis of both equirectangular images using the extrinsic transform obtained by the calibration.

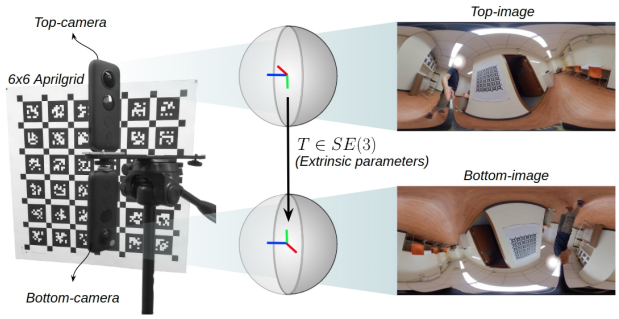

Fig. 3: Our 360° stereo system composed of two Insta360® ONE X cameras. In order to align both equirectangular images (top and bottom), the extrinsic parameters between the cameras is needed. This transformation is obtained by stereo calibration.

B. Metrics

We have evaluated both depth and disparity results using MAE and RMSE. The depth error is prone to increasing significantly based on the non-linear relationship between depth and disparity as in (1), which does not provide informative evaluation. Therefore, we crop out 5% of largely distorted depth map from the top and bottom, respectively.

C. Experimental Setting

Our model is trained from scratch with Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) solver for 400 epochs with an initial learning rate of 0.001 and fine-tuned with a learning rate of 0.0001 for 100 epochs on MP3D. For SF3D, we follow the same setting as MP3D but with 50 epochs using pre-trained model from MP3D. The entire implementation is based on the PyTorch framework.

D. Overall Performance

In Table I, we show results on MP3D and SF3D with comparisons to state-of-the-art stereo depth estimation approaches, including the conventional methods (ASW [3], Binocular [4] and KIM's [5]) and deep learning-based models (PSMNet [1] and GCNet [2]). Our method achieves significant improvement for both the disparity and depth performances, since other methods do not consider distortion introduced in 360° images. These results demonstrate the effectiveness of our designs for 360° images, including polar angle and LCV modules. In addition, compared to the baseline PSMNet model, our method only introduces a slight overhead in runtime, while our model outperforms PSMNet by a large margin.

E. Ablation Study

We present an ablation study in Table II on MP3D for depth estimation to validate the effectiveness of each component in the proposed framework. Comparing ID 1 with

TABLE I: Experimental results of the proposed method on MP3D and SF3D compared with other approaches including deep learning-based networks and conventional algorithms.(↓ represents the lower the better.)

	MP3D					SF3D				
	Disparity		Depth		Time	Disparity		Depth		Time
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓		MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	
Binocular [4]	0.7206	2.507	0.1368	0.5399	0.6333	0.3204	1.5494	0.0897	0.4496	0.6333
KIM's [5]	0.8175	2.2956	0.2191	0.6955	1.8507	2.5327	4.39	0.1163	0.3972	1.8507
ASW [3]	0.4410	1.648	0.1427	0.5193	7.5min	0.2155	0.7754	0.0779	0.2628	7.5 min
GCNet [2]	0.486	1.4283	0.0969	0.2953	1.54s	0.1877	0.4971	0.0592	0.1361	1.57s
PSMNet [1]	0.3139	1.049	0.0946	0.2838	**0.50s**	0.1292	0.4053	0.0418	0.1068	**0.51s**
360SD-Net (Ours)	**0.1447**	**0.6930**	**0.0593**	**0.2182**	0.572s	**0.1034**	**0.3691**	**0.0335**	**0.0914**	0.55s

TABLE II: Ablation study for depth estimation on MP3D. The first row **bs** is considered as the baseline in this study, which uses a fixed step-size vertical pixel shifting.. Different components are denoted as: (**Pc**) Polar angle with early fusion; (**Pb**) Polar angle with late fusion; (**ASPP**) ASPP module; (**LCV**) Learnable Cost Volume; (**repli**) LCV with replicate padding.

ID	Ablation Study	Depth RMSE ↓
1	bs	0.2765
2	bs + Pc (Table III Coordinate ID2)	0.2501
3	bs + Pb	0.2494 (+9.8%)
4	bs + Pb + ASPP	0.2462 (+10.9%)
5	LCV (Table III Step Size ID3)	0.2464
6	LCV (repli)	0.2409 (+12.9%)
7	(Ours) bs + Pb + ASPP + LCV (repli)	**0.2182** (+21.1%)

TABLE III: Ablation study of different coordinate information added and different initial step-size of LCV for depth estimation on MP3D.

ID	Coordinate	Depth RMSE ↓	Step Size	Depth RMSE ↓
1	horizontal angle	0.2583	1°	0.2611
2	polar angle	**0.2501**	1/2 °	0.2559
3	radius	0.2541	1/3 °	**0.2464**
4	arc-length	0.2516	1/4 °	0.2503
5	area	0.2513	-	-

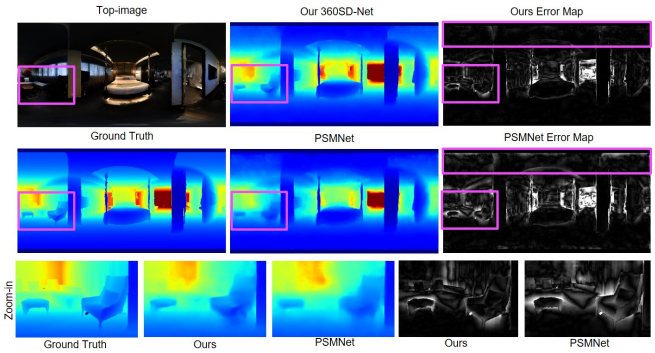

Fig. 4: Qualitative depth map and error map comparison between 360SD-Net (Ours) and PSMNet. Our depth map shows sharper and clearer details in both close and distant regions. For the zoom-in views, the armchair and table present a notable geometry structure compared to the one from PSMNet. Our error map also shows higher accuracy in regions with higher distortion and object boundaries.

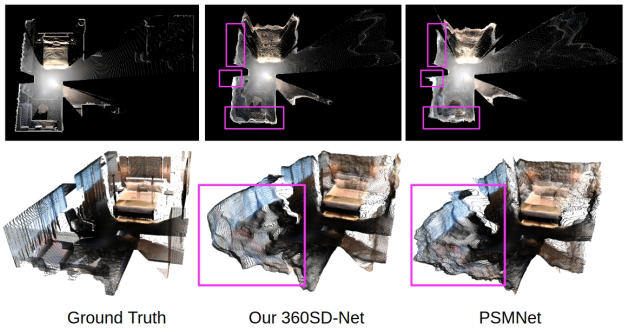

Fig. 5: Qualitative point cloud comparison between 360SD-Net (Ours) and the PSMNet. Our model shows a better geometry estimation with less distortion and a more accurate structure.

5, it shows the effectiveness of LCV, while ID 2 shows the benefits from the polar angle. The other rows gradually show the improvement of adding other designs such as ASPP and replicated-padding. With the combination of ID 4 and 6, we form our final network that achieves the best performance.

Detailed ablation studies on polar angle and LCV are shown in Table III, which compares different geometry measurements from spherical projection and different initial step-sizes in degree applied in LCV. Through comparing various geometry measurements, including area, arc-length, radius, and horizontal angle, using the polar angle performs the best in dealing with distortion. Regarding initial step-size in LCV, we demonstrate empirically that the performance increases when the initial step-size value decreases. The improvements saturate at $\frac{1}{3}°$, which is chosen to be our initial step-size value.

F. Qualitative Results

We present qualitative results in depth maps and point clouds, mainly compared with PSMNet [1] based on its good performance in Table I. As shown in Fig. 4 and Fig. 5, our model results in sharper depth maps and our projected point clouds are able to reconstruct scenes more accurately in comparison to PSMNet. Furthermore, Fig. 6 shows more qualitative results on both datasets of our model.

G. Qualitative Results for Real-World Images

To show the generalization of our model (trained on the synthetic MP3D dataset) on real-world scenes, we take still

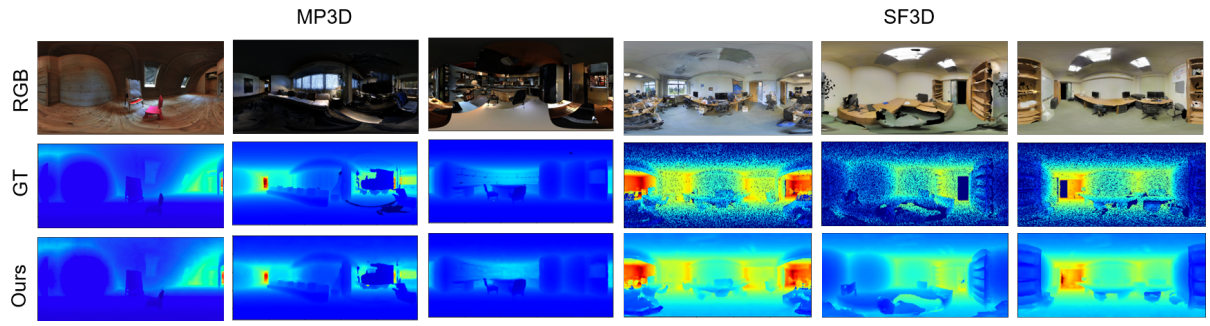

Fig. 6: More qualitative results for depth map on MP3D and SF3D. For MP3D, our estimated depth maps preserve object and surface details with results similar to GT. For SF3D, our model outputs dense depth maps of high accuracy, with training on sparse GT.

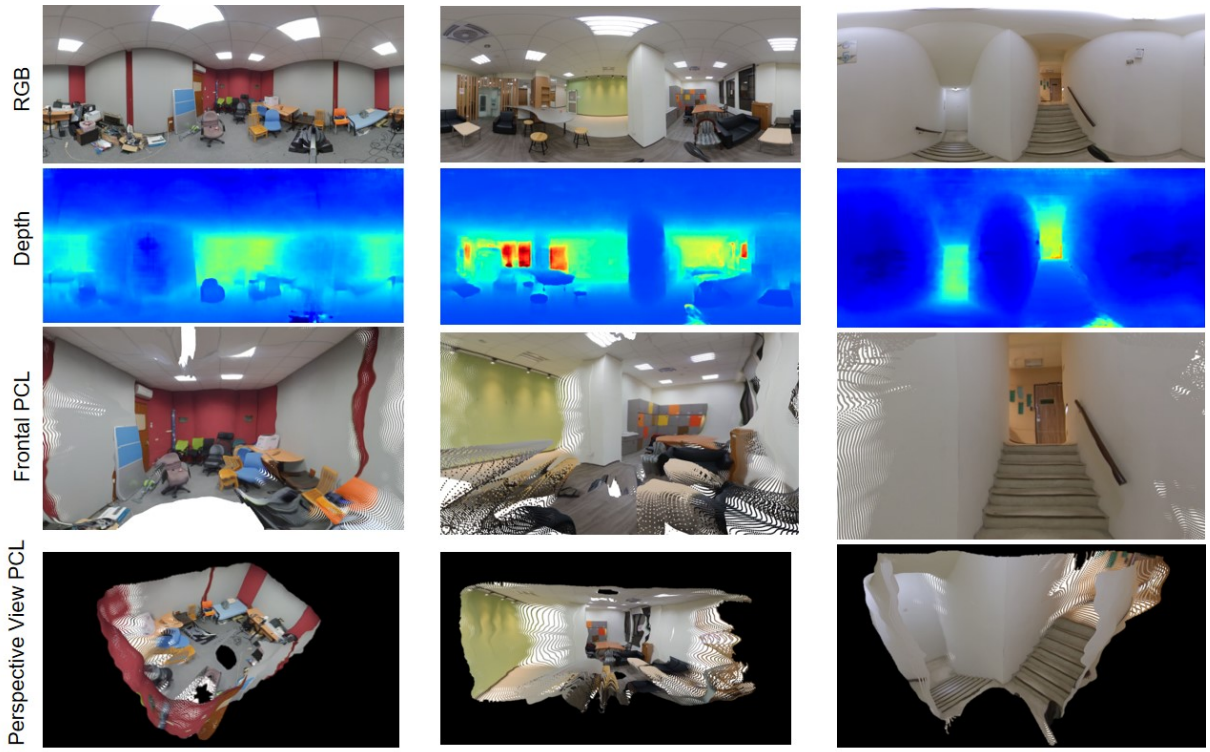

Fig. 7: Qualitative results on real scenes using two Insta360® ONE X cameras in a top-bottom configuration. The furniture can be clearly seen in the depth maps and also well reconstructed in the point clouds.

and moving images with a pair of well-known consumer-level 360° cameras. In order to reduce the domain gap, we apply our model on these real-world images using gray-scale. In Fig. 7, we show the results in depth maps, frontal view, and perspective view of point clouds with their regarded RGB images. The details of objects and room layouts are elegantly reconstructed, which shows great compatibility of our network between synthetic and real-world scenes. Moreover, our model produces promising depth maps for handheld videos (refer to supplementary video for more results).

V. Conclusions

In this paper, we introduce the first end-to-end trainable deep network, namely 360SD-Net, for depth estimation directly on 360° stereo images via designing a series of improvements over existing methods. In experiments, we show state-of-the-art performance on our collected synthetic datasets with extensive ablation study that validates proposed modules, including the usage of polar angle and learnable cost volume design. Finally, we test on real-world scenes and present promising results with the model trained on the pure synthetic data to show the generalization and compatibility of our presented network.

Acknowledgement. This project is supported by the National Center for High-performance Computing, MOST Joint Research Center for AI Technology and All Vista Healthcare with program MOST 108-2634-F-007-006 and MOST 109-2634-F-007-016.

REFERENCES

[1] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[2] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[3] K. J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, 2006.

[4] S. Li, "Binocular spherical stereo," *IEEE Transactions on intelligent transportation systems*, vol. 9, no. 4, pp. 589–600, 2008.

[5] H. Kim and A. Hilton, "3d scene reconstruction from multiple spherical stereo pairs," *International Journal of Computer Vision*, vol. 104, 08 2013.

[6] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[7] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese, "Joint 2D-3D-Semantic Data for Indoor Scene Understanding," *ArXiv e-prints*, Feb. 2017.

[8] C. Banz, S. Hesselbarth, H. Flatt, H. Blume, and P. Pirsch, "Real-time stereo vision system using semi-global matching disparity estimation: Architecture and FPGA-implementation," in *2010 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation*, 2010.

[9] G. S. Hong and B. G. Kim, "A local stereo matching algorithm based on weighted guided image filtering for improving the generation of depth range images," *Displays*, 2017.

[10] R. A. Hamzah, M. S. Hamid, A. F. Kadmin, S. F. A. Gani, S. Salam, and T. M. Wook, "Accurate Disparity Map Estimation Based on Edge-preserving Filter," in *2018 International Conference on Smart Computing and Electronic Enterprise, ICSCEE 2018*. IEEE, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8538360/

[11] S. B. Kang and R. Szeliski, "3-d scene data recovery using omnidirectional multibaseline stereo," *International journal of computer vision*, vol. 25, no. 2, pp. 167–183, 1997.

[12] S. Im, H. Ha, F. Rameau, H.-G. Jeon, G. Choe, and I. S. Kweon, "All-around depth from small motion with a spherical panoramic camera," in *European Conference on Computer Vision*. Springer, 2016, pp. 156–172.

[13] G. Koch, "Siamese neural networks for one-shot image recognition," 2015.

[14] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.

[15] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695–5703.

[16] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," *CoRR*, vol. abs/1801.10130, 2018. [Online]. Available: http://arxiv.org/abs/1801.10130

[17] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning so (3) equivariant representations with spherical cnns," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 52–68.

[18] Y. Su and K. Grauman, "Kernel transformer networks for compact spherical convolution," *CoRR*, vol. abs/1812.03115, 2018. [Online]. Available: http://arxiv.org/abs/1812.03115

[19] ——, "Flat2sphere: Learning spherical convolution for fast features from 360° imagery," *CoRR*, vol. abs/1708.00919, 2017. [Online]. Available: http://arxiv.org/abs/1708.00919

[20] C. Zou, A. Colburn, Q. Shan, and D. Hoiem, "Layoutnet: Reconstructing the 3d room layout from a single rgb image," in *CVPR*, 2018.

[21] S.-T. Yang, F.-E. Wang, C.-H. Peng, P. Wonka, M. Sun, and H.-K. Chu, "Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3363–3372.

[22] C. Sun, C.-W. Hsiao, M. Sun, and H.-T. Chen, "Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360° videos," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[24] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360° videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[25] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[26] F. Wang, H. Hu, H. Cheng, J. Lin, S. Yang, M. Shih, H. Chu, and M. Sun, "Self-supervised learning of depth and camera motion from 360° videos," *CoRR*, vol. abs/1811.05304, 2018. [Online]. Available: http://arxiv.org/abs/1811.05304

[27] G. Payen de La Garanderie, A. Atapour Abarghouei, and T. P. Breckon, "Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360° panoramic imagery," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[28] C. Won, J. Ryu, and J. Lim, "Sweepnet: Wide-baseline omnidirectional depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 6073–6079.

[29] M. Eder, P. Moulon, and L. Guan, "Pano popups: Indoor 3d reconstruction with a plane-aware network," *2019 International Conference on 3D Vision (3DV)*, Sep 2019. [Online]. Available: http://dx.doi.org/10.1109/3DV.2019.00018

[30] N. Zioulis, A. Karakottas, D. Zarpalas, F. Alvarez, and P. Daras, "Spherical view synthesis for self-supervised 360° depth estimation," *2019 International Conference on 3D Vision (3DV)*, Sep 2019. [Online]. Available: http://dx.doi.org/10.1109/3DV.2019.00081

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[32] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, "MINOS: Multimodal indoor simulator for navigation in complex environments," *arXiv:1712.03931*, 2017.

[33] F.-E. Wang, H.-N. Hu, H.-T. Cheng, J.-T. Lin, S.-T. Yang, M.-L. Shih, H.-K. Chu, and M. Sun, "Self-supervised learning of depth and camera motion from 360°videos," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 53–68.

[34] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 8, pp. 1335–1340, 2006.