# IRS: A Large Synthetic Indoor Robotics Stereo Dataset for Disparity and Surface Normal Estimation

Qiang Wang[1,*], Shizhen Zheng[1,*], Qingsong Yan[2,*], Fei Deng[2], Kaiyong Zhao[1,†], Xiaowen Chu[1,†]

***Abstract*— Indoor robotics localization, navigation and interaction heavily rely on scene understanding and reconstruction. Compared to monocular vision which usually does not explicitly introduce any geometrical constraint, stereo vision based schemes are more promising and robust to produce accurate geometrical information, such as surface normal and depth/disparity. Besides, deep learning models trained with large-scale datasets have shown their superior performance in many stereo vision tasks. However, existing stereo datasets rarely contain the high-quality surface normal and disparity ground truth, which hardly satisfy the demand of training a prospective deep model for indoor scenes.**

**To this end, we introduce a large-scale synthetic indoor robotics stereo (IRS) dataset with over 100K stereo RGB images and high-quality surface normal and disparity maps. Leveraging the advanced rendering techniques of our customized rendering engine, the dataset is considerably close to the real-world captured images and covers several visual effects, such as brightness changes, light reflection/transmission, lens flare, vivid shadow, etc. We compare the data distribution of IRS with existing stereo datasets to illustrate the typical visual attributes of indoor scenes. In addition, we present a new deep model DispNormNet to simultaneously infer surface normal and disparity from stereo images. Compared to existing models trained on other datasets, DispNormNet trained with IRS produces much better estimation of surface normal and disparity for indoor scenes.**

## I. Introduction

Indoor scene understanding and reconstruction are central to many robotics applications, such as robot localization, navigation, and interaction. Despite the attractive cost and availability, monocular vision does not explicitly introduce any geometrical constraint. On the contrary, stereo vision leverages the advantage of cross-reference between the left and the right view, and usually shows greater performance and robustness in geometric information inference, such as surface normal and disparity/depth estimation. Recent advances [1], [2], [3], [4], [5], [6] in these vision tasks have shown that deep neural network (DNN) can significantly improve the estimation quality. However, the success of DNN requires large scale and high-quality labelled datasets, which are still lacking in stereo vision studies.

Surface normal and disparity/depth are two core information in 3D geometry understanding since they can determine the position and orientation of an object in the space. In addition, they also have strong knowledge relation. On the one hand, surface normal is determined by local surface tangent plane of neighboring 3D points, which can be calculated from depth; on the other hand, the orientation of the plane constructed by the depth is constrained by the surface normal. This knowledge has been used in [7], [8], [9] to jointly optimize the quality of binocular vision.

Existing studies [10], [11], [12], [13] proposed stereo datasets collected by real sensing hardware, which contributed a lot to stereo vision research. However, they typically have only a small number of samples and lack complete and dense ground-truth of surface normal and disparity. Recent work in [14], [15], [16] leveraged synthetic technology to generate sufficiently large data volume for DNN training. However, there are two main drawbacks of them. First, few of stereo vision datasets contains both high-quality disparity/depth and surface normal ground truth. Second, due to the limitation of the rendering systems, their stereo RGB images are usually noisy and not realistic enough in terms of brightness variation, light reflection/transmission, indirect shadows, bloom, lens flare, etc. It has been shown that the existing state-of-the-art deep models [4], [15] trained on these synthetic datasets did not work well on the complicated real-world indoor scenes.

To this end, we propose IRS, a large scale synthetic stereo dataset for indoor robotics applications, which is generated by a customized advanced rendering engine. We also conduct quantitative analysis and deep model training experiments to verify the effectiveness of learning indoor geometrical information from IRS. Our contributions are summarized as follows:

- We present a large synthetic indoor robotics stereo dataset, namely IRS, generated by a customized version of Unreal Engine (UE4) with originally implemented plug-ins. Our dataset contains over 100K of stereo RGB pairs as well as their complete surface normal and disparity ground truth. With the advanced rendering techniques provided by UE4, the vision attributes of the real-world physical environment can be well simulated, including light reflection/transmission, bloom, lens flare, etc.
- We conduct quantitative analysis to compare IRS with some existing stereo datasets. We show that IRS covers common visual attributes of indoor scenes, including lightness variation and camera vision range changes.
- We present a new deep model, DispNormNet[1], to jointly

*Authors have contributed equally.

†Corresponding authors.

[1]Department of Computer Science, Hong Kong Baptist University, {qiangwang,szzheng,kyzhao,chxw@comp.hkbu.edu.hk}

[2]School of Geodesy and Geomatics, Wuhan University, {yanqs_whu@whu.edu.cn, fdeng@sgg.whu.edu.cn}

[1]All experimental settings and source codes can be found at GitHub: https://github.com/HKBU-HPML/IRS

TABLE I: The Comparison of Recent Stereo Datasets and Our IRS

| **Dataset** | MiddleBury[12] | KITTI2012[13] | KITTI2015[17] | Sintel[14] | Apollo [18] | Scene Flow[15] | IRS(ours) |
|---|---|---|---|---|---|---|---|
| **Synthetic(S) / Natural(N)** | N | N | N | S | N | S | S |
| **Scene** | Lab | Road | Road | Outdoor | Road | Outdoor | Indoor |
| **Resolution** | 2960x1942 | 1226x370 | 1242x375 | 1024x436 | 3130x960 | 960x540 | 960x540 |
| **Training/Testing Data Size** | 23/10 | 194/194 | 200/200 | 1064/0 | 22390/0 | 35454/4370 | 84946/15079 |
| **Density of Ground Truth** | 100% | $\leq$ 30% | $\leq$ 30% | 100% | 100% | 100% | 100% |
| **Number of Texture Types** | 1 | 1 | 1 | 1 | 1 | Multiple | Multiple |
| **Surface Normal** | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Textureless Region** | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\times$ | $\surd$ |
| **Reflection & Light** | $\surd$ | $\times$ | $\times$ | $\times$ | $\surd$ | $\times$ | $\surd$ |

predict the surface normal and disparity for indoor scene stereo images. Compared to existing models trained on Scene Flow, another large scale stereo datasets, DispNormNet yields competitive performance on both synthetic and real-world indoor stereo data and shows decent robustness to intensive brightness changes and light reflection/transmission of glass and mirrors.

The rest of the paper is organized as follows. We introduce the existing studies on stereo vision datasets in Section II. Then we present the methodology and implementation of generating IRS dataset in Section III. Section IV presents a quantitative comparison between IRS and existing stereo datasets. In Section V, we present our performance evaluation to illustrate the effectiveness of IRS on indoor synthetic and real-world stereo data. We finally conclude the paper in Section VI.

## II. Related Work

Indoor scene understanding is vital to many robotics application. While monocular vision mainly focuses on core scene understanding tasks such as object detection [19], [20] and semantics segmentation [21], [22], stereo vision is popular to infer the spatial geometric information, including surface normal and depth/disparity, which play a significant role in the fields of autonomous driving and indoor robots. Compared to monocular disparity estimation and normal estimation which rely on the prior information of the scene and lack geometric constraints, stereo vision can combine the prior information and geometry information together to give better estimated results [7], [8], [9].

Before the popularity of DNN models on solving big data training tasks, the main usage of the datasets is to evaluate different algorithms, which means that the data complexity and variety is more important than the data volume. MiddleBury is a well-known and frequently updated real world stereo dataset [23], [10], [11], [24], [12], which takes the resolution, brightness, exposure and many other uncommon factors into consideration to improve the complexities of the dataset and provides dense per-pixel disparities of the scenes.

With the rapid development of deep learning methods, traditional datasets can no longer meet the demand of training a deep model of decent robustness and generalization. Datasets with larger volume and more complicated distributions have received more attention.

Sintel [14] is a synthetic dataset based on open source 3D movies. It uses Blender 3D engine to render the scene and obtain corresponding depth information and fully considers the influence of various factors, like motion blur and defocus blur. It provides 1,064 stereo images with high-quality disparity maps and has been used to train effective networks on realistic data in [15], [2], [1].

KITTI is a natural dataset for autonomous driving, which provides 394 pairs to train and 394 pairs to test. KITTI2012 [13] released the real world captured stereo images on roads as well as their disparity maps of high sparsity obtained by Velodyne HDL-64E. KITTI2015 [17] extends the dataset by modeling cars, which may exist some fitting errors.

Scene Flow [15] has over 39,000 pairs to train the network, which is based on the 3D model provided by ShapeNet [25] and the texture from Flicker. Unlike other synthetic datasets that based on the time-consuming 3D engine, the data of Scene Flow is constructed by random selection of scene backgrounds, 3d objects, and the textures on those objects.

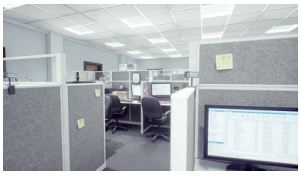
Apollo Stereo [18], like KITTI, is also a dataset for autonomous driving, but it uses two VUX-1HAs to get more denser depth information, providing a total of 22,390 pairs for training. Its ground truth is acquired by accumulating 3D point clouds from Lidar and fitting 3D CAD models to individually moving cars.

Currently, stereo datasets are either object-centric, such as Sintel and Scene Flow, where objects usually appear in the center of the image; or for autonomous driving, of which the main scenes are roads. With the development of indoor robots, stereo vision solutions for indoor scenes is becoming increasingly important. This paper presents a dataset for indoor scenes, and the relevant information is shown in Table I. Our IRS is produced by an advanced rendering engine and close to the real world captured pictures. It also provides high-quality and dense labelled ground truth for surface normal and disparity.

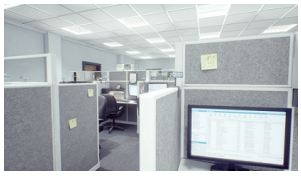
## III. Data Generation

We developed a customized version of the Unreal Engine 4 (UE4, version 4.21) to generate stereo images and the ground truth of disparity and surface normal for different virtual indoor scenes. With the advanced rendering techniques, such as ambient occlusion, diffuse inter-reflection, etc., UE4 provides high simulation effects close to the real world captured images in terms of light and shadow effects, brightness variation, bloom and len flare, and excellent surface materials. It is also equipped with the convenient scene editor and C++ API for self-customized plug-in development. We modified the

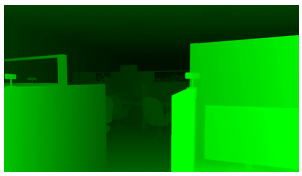
pipeline of UE4's internal rendering engine and customized a plug-in[2] to produce stereo RGB images as well as their corresponding disparity maps and surface normal maps. Fig. 1 shows one example. We used deferred rendering to produce stereo RGB images. In our rendering, we cover many factors in real-world indoor scenes, including highlights, light colors, over-exposure, shadows, dark environments, specular reflections, metal surfaces, noise, etc., as illustrated in Fig. 2. Then we used the contents in GBuffer, which are generated during the rendering process, to make up disparity maps and normal maps. The high-quality dense ground truth of surface normal and disparity maps are critical to train a prospective machine learning model.



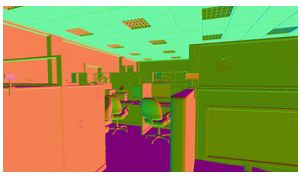
(a) Left RGB Image



(b) Right RGB Image



(c) Left Disparity Map



(d) Left Surface Normal Map

Fig. 1: Data samples generated by our customized UE4.

Table II concludes the data distribution of IRS. IRS contains more than 100,000 pairs of 960x540 resolution stereo images (84,946 training and 15,079 testing), covering four indoor scene types and 70 different scene instances. The scenes are all enclosed indoor layout, and some of them even have a visible distance longer than 20 meters. We place over 2,091 identical objects of different types in the constructed space. We also consider different cases of brightness and light behaviors commonly happening in the indoor environments.

## IV. Quantitative Analysis

In this section, we quantitatively compare our IRS with those existing stereo datasets in terms of the distribution of normal, disparity and brightness.

### A. Normal Analysis

As IRS is the only stereo dataset that provides normal information, we only count the distribution of the normal of IRS. We first convert the three-dimensional normal vector $n = (n_x, n_y, n_z)$ into two dimensional $n_{angle} = (\alpha = atan2(n_y, n_x), \beta = atan2(n_z, \sqrt{n_x^2 + n_y^2})$, where $alpha$ represents the angle of $n$ in the $x-y$ plane with the range $[0°, 360°]$, and $beta$ represents the angle between $n$ and the $x-y$ plane, ranging from $[-90°, 90°]$. Notice that $[0°, 90°]$ is an invisible area for the camera.

[2]The plug-in will be open-sourced once available.

(a) Home

(b) Office

(c) Restaurant

(d) Dark

(e) Normal Light

(f) Over Exposure

(g) Glass

(h) Mirror

(i) Metal

Fig. 2: The first row shows different indoor scenes. The second row shows different lightness environments. The third row shows the light transmission/reflection effects of different materials.

TABLE II: The indoor scene types and visual attributes covered by our IRS.

| Rendering Variable | Options |
|---|---|
| indoor scene | home(30995), office(41987), restaurant(20969), store(6347) |
| object | desk, chair, sofa, glass, mirror, bed, bedside table, lamp, wardrobe, etc. |
| brightness | over-exposure(>1300), darkness(>1700) |
| light behavior | bloom(>1700), lens flare(>1700), glass transmission(>3600), mirror reflection(>3600) |

First, we average all the distributions of the normal vectors in each sample of IRS. Then, we transform the result by $log$ function, as shown in Fig. 3. As the images of IRS is captured indoors, the normal vector is mostly located around four directions, $(0°, 0°)$, $(90°, 0°)$, $(180°, 0°)$, $(270°, 0°)$, which respectively correspond to the left wall, the floor, the right wall, and the ceiling of the scene. There are also a large number of normal vectors with $\beta = -90°$, which corresponds to the wall facing the camera.

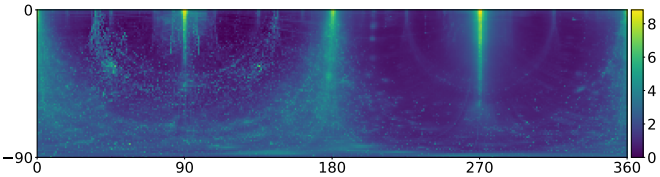


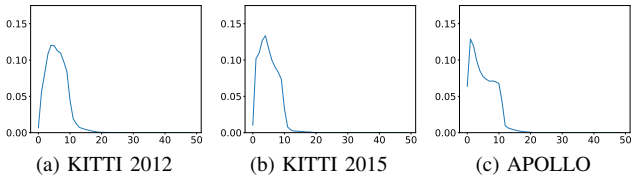

Fig. 3: Normal Distribution

### B. Disparity Analysis

To quantitatively evaluate and compare the disparity distributions of different datasets, we first divide all the disparity

values by the image width. Then we enlarge those divided results by 200× for better visualization.

We analyze six different datasets and draw their disparity distributions in Fig. 4, where the $x$ axis represents the preprocessed disparity values and the $y$ axis represents the percentage. As most of the values are less than 50, we only present the range of 0-50. In Fig. 4, (a)-(c) are the natural datasets and (d)-(f) are synthetic datasets. Sintel and Scene Flow are two of the most commonly used datasets, but there are some obvious troughs in their distributions, which are quite different from the natural datasets. However, our proposed IRS has no such problem and performs a similar distribution to those of two natural datasets.



(a) KITTI 2012 (b) KITTI 2015 (c) APOLLO

(d) SINTEL (e) SCENE FLOW (f) OURS

Fig. 4: Disparity Distribution.

### C. Brightness Analysis

Considering that DNNs mainly learn disparity through texture information, we also analyze the brightness distribution of matched pixels.

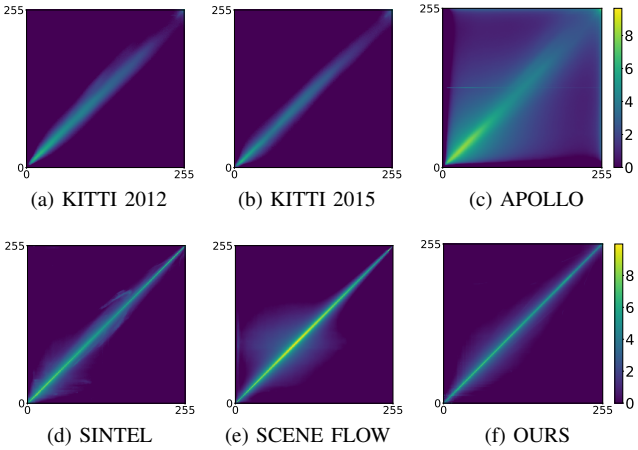
We first convert all the RGB images into gray-scale. The results are shown in Fig. 5, where the $x$ axis is the brightness of the left image, the $y$ axis is the brightness of the right image, and the value represents the logarithm of the average number of matched pixels on each pair. The distribution of matched pixels in each dataset roughly spreads along a line ($y = x$); in natural datasets, such as KITTI and Apollo, the distribution is relatively discrete due to the uncontrollable external environments; but in the synthetic datasets, most of those matched pixels are close to $y = x$, which means that few brightness changes happen between left and right images. To this end, synthetic datasets usually need data enhancement to fill the gap with the natural datasets.

In addition, in the natural datasets, overexposure is a very common phenomenon. In Fig. 6, the three natural datasets, KITTI 2012, KITTI 2015 and Apollo, have a large number of matched pixels at $(255, 255)$; but the existing synthetic datasets, like Sintel and Scene Flow, do not have enough overexposure pixels. In IRS, we purposely create enough overexposed pixels to simulate the real scenarios.

## V. Performance Evaluation

### A. Experimental Setup

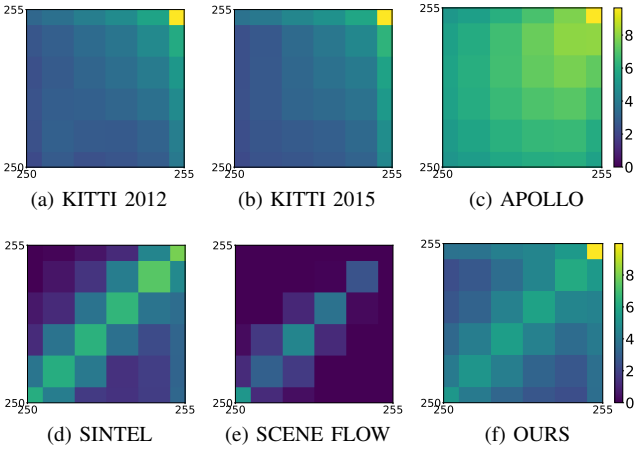
*1) Disparity Estimation:* To validate the effectiveness of our proposed IRS for indoor disparity estimation, we used



(a) KITTI 2012 (b) KITTI 2015 (c) APOLLO

(d) SINTEL (e) SCENE FLOW (f) OURS

Fig. 5: Brightness Distribution.



(a) KITTI 2012 (b) KITTI 2015 (c) APOLLO

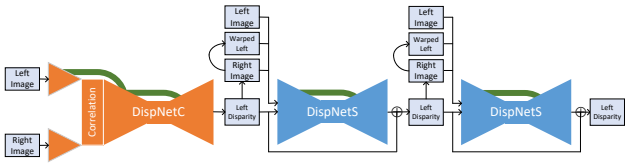
(d) SINTEL (e) SCENE FLOW (f) OURS

Fig. 6: Over-Exposure Pixels Distribution.

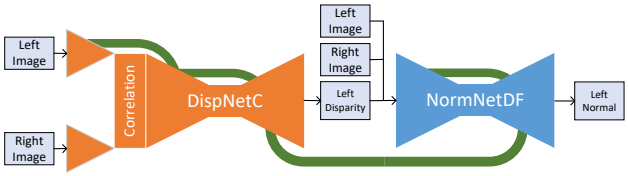
different training datasets to train the same DNN architecture for constructing different DNN models. Then we applied these models on different testing datasets and obtained the errors. We used the endpoint error (EPE) as error measure in all the cases. We chose DispNet-CSS[2] from one of the state-of-the-art models as our basic DNN architecture, as shown in Fig. 7a.

We trained two models, DN-CSS-IRS and DN-CSS-SF, respectively on the IRS training data and the Scene Flow training data using the DispNet-CSS architecture. Then we applied them to different testing datasets listed in Table III.

*2) Normal Estimation:* As for surface normal estimation from stereo images, we designed a novel deep model, DispNormNet, to jointly train the surface normal and disparity meanwhile. Fig. 7b demonstrates the structure of DispNormNet, which is comprised of two modules, DispNetC and NormNetDF. DispNetC is identical to that in [15] and produces the disparity map. NormNetDF produces the normal map and is similar to DispNetS except two modifications. First, the final convolution layer outputs three channels instead of one, as each pixel normal has three dimension $(x, y, z)$. Second, we concatenate the deconvolution features




(a) DispNet-CSS




(b) DispNormNet

Fig. 7: DispNet-CSS[2] and our DispNormNet.

of DispNetC to that of NormNetDF with the same feature map size in turn. "DF" indicates disparity feature fusion, which we found important to produce accurate surface normal maps.

To compare the predicted normal to the ground truth, we first calculate the angle between them and take the mean value and the median value. We then compute the fraction of pixels of which the angle error is less than $t$, where $t = 11.25°, 22.5°, 30°$, as adopted in [26].

*3) Training Implementation:* We implemented DispNet-CSS and DispNormNet using PyTorch. All the models were end-to-end trained with Adam ($\beta_1 = 0.9, \beta_2 = 0.999$). We performed color normalization with the mean and variation of ImageNet [27] for data preprocessing. During training, images were randomly cropped to size $H = 384$ and $W = 768$. The batch size was set to 16 on four Nvidia Titan X Pascal GPUs (each of 4). To ensure the fairness of model training, we trained the model for 70000 iterations on the selected training dataset. The learning rate was initialized as $10^{-4}$ and decreased by half every 25000 iterations.

### B. Experimental Results

*1) Performance on Indoor Scene Disparity:* We first explore the model accuracy on the testing samples of two synthetic datasets. Table III lists the EPEs of different cases. First, it is obvious and reasonable that DN-CSS-SF and DN-CSS-IRS perform better than each other on their own testing samples due to the similar RGB data distribution. However, DN-CSS-IRS even yield good generalization with a 4.791 average EPE on Scene Flow testing data, while it is not the case for DN-CSS-SF. We make a deeper analysis by comparing the EPEs of several IRS testing data partition with different visual attributes, including over-exposure, darkness, len flare and glass/mirror reflection. It is observed that DN-CSS-SF suffers a lot from the former three attributes, which are caused by camera optical characteristics and environment brightness, especially darkness. Besides, it is revealed that glass/mirror reflection and transmission is the most challenging visual factor for indoor scene understanding, as indicated by high EPEs of both models on those type of images.

Fig. 8 illustrates some typical examples of disparity maps predicted by DN-CSS-SF and DN-CSS-IRS on our IRS dataset. The first row shows a RGB image with lens flare. DN-CSS-SF mistakes the lens flare patch as background and predicts very small disparity values for those pixels, while DN-CSS-IRS performs more robust results. The second row shows a wide view range case, which also confuses DN-CSS-SF in those far regions. The third row shows an over-exposure case, which misleads DN-CSS-SF into producing unreasonable results for those regions. The last row shows an image containing glass and mirrors. Since the data of Scene Flow contains various of texture types, which tends to teach the network to learn disparity by feature matching, reflection and perspective of the environment can easily confuse DN-CSS-SF. However, DN-CSS-IRS trained on IRS can learn this kind of knowledge and produce much better results.

TABLE III: The Disparity Errors of Three Models on Different Datasets. $\underline{X}$ indicates the best.

| Testing Data | DN-CSS-SF | DN-CSS-IRS | DispNormNet |
|---|---|---|---|
| Scene Flow | <u>1.46</u> | 4.791 | 5.969 |
| IRS | 7.18 | <u>2.33</u> | 2.37 |
| IRS(Over Exposure) | 4.578 | 2.343 | <u>2.283</u> |
| IRS(Darkness) | 10.908 | <u>1.892</u> | 2.096 |
| IRS(Lens Flare) | 3.647 | <u>1.772</u> | 1.786 |
| IRS(Glass/Mirror) | 4.902 | 3.529 | <u>3.392</u> |

*2) Performance on Indoor Scene Surface Normal:* Table IV concludes the prediction accuracy of normal maps of DispNormNet. It reveals that DispNormNet can predict considerably accurate normal maps with a mean angle error of 17.33°. Even 50% of pixels have low errors of no more than 10°. The error distribution also guarantees that over 76% of predictions are within a fault tolerance range of [0, 22.5°].

TABLE IV: The Normal Errors of DispNormNet on IRS.

| | mean | median | <11.25° | <22.5° | <30° |
|---|---|---|---|---|---|
| DispNormNet | 17.33° | 9.97° | 54.9% | 76.1% | 82.4% |

Fig. 9 demonstrates some examples of normal maps predicted by DispNormNet. It is observed that the predicted results are remarkably close to the ground truth in terms of object shape and smooth plane. However, some small objects in the image are still messed and cannot be recognized. We leave it as our future work of developing a better network training scheme.

*3) Qualitative Results on Real World Images:* We also evaluate the performance of DispNormNet on the real world data. We use Intel RealSense D435i stereo camera to capture some indoor images of offices and canteens and then apply DispNormNet to predict their disparity and surface normal maps. Fig. 10 illustrates some examples of the captured left images. DispNormNet yields superior robustness to the complicated lightness environment and irregular shadows, as indicated by the first row. Then the second example contains the mirror-surface ground which reflects the light. It is surprising that DispNormNet assigns those regions with a smooth disparity and normal map, which is reasonable for a plane. The third example shows a wall with rich texture and a long aisle. DispNormNet can still notice the flatness


Fig. 8: Disparity Prediction Results on Synthetic Images.

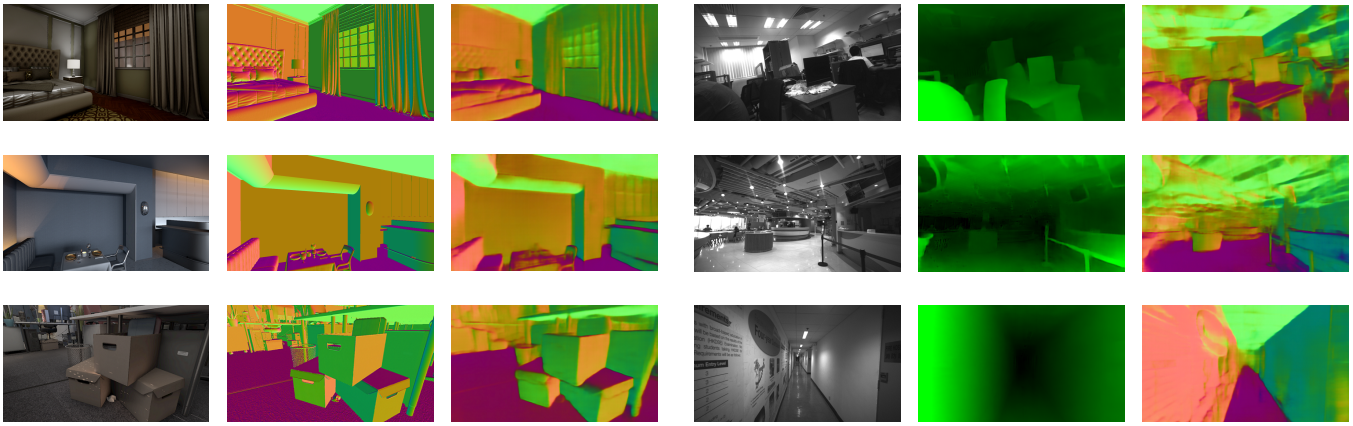



Fig. 9: Surface Normal Prediction Results on Synthetic Images.




Fig. 10: Disparity Prediction Results of DispNormNet on Real World Images.

of walls, floors and ceiling, and produce satisfying disparity and normal maps.

## VI. Conclusion

In this paper, we propose IRS, a large-scale synthetic stereo dataset for indoor robotics targeted at disparity and surface normal estimation. IRS covers a wide range of indoor scenes, including office, home, restaurant and store, and are remarkably close to the real world captured images in the aspects of brightness changes, light reflection/transmission, lens flare, etc. To illustrate the usage and functionality of IRS, we first compare the data distribution of IRS to other stereo datasets to analyze the classical visual attributes of indoor scenes. We then develop a deep model, DispNormNet, to predict both surface normal and disparity from stereo images. Our experimental results indicate that DispNormNet trained with IRS produces much better results than existing models trained with other stereo datasets. DispNormNet also demonstrates its great potential for surface normal and disparity estimation of real world indoor scenarios.

## REFERENCES


[1] R. Atienza, "Fast disparity estimation using dense networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 3207–3212.

[2] E. Ilg, T. Saikia, M. Keuper, and T. Brox, "Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[3] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[6] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo-stereo matching with slanted support windows." in *Bmvc*, vol. 11, 2011, pp. 1–11.

[8] S. Zhang, W. Xie, G. Zhang, H. Bao, and M. Kaess, "Robust stereo matching with surface normal prediction," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2540–2547.

[9] D. Scharstein, T. Taniai, and S. N. Sinha, "Semi-global stereo matching with surface orientation priors," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 215–224.

[10] D. Scharstein and R. Szeliski, "High accuracy stereo depth maps using structured light," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.

[11] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[12] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42.

[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[14] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European conference on computer vision*. Springer, 2012, pp. 611–625.

[15] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[16] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser, "Physically-based rendering for indoor scene understanding using convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] M. Menze, C. Heipke, and A. Geiger, "Joint 3d estimation of vehicles and scene flow," in *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[18] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 954–960.

[19] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[21] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, "Single-shot object detection with enriched semantics," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[23] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002.

[24] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[25] 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.

[26] J. Zeng, Y. Tong, Y. Huang, Q. Yan, W. Sun, J. Chen, and Y. Wang, "Deep surface normal estimation with hierarchical rgb-d fusion," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.