

DIODE: A Dense Indoor and Outdoor DEpth Dataset

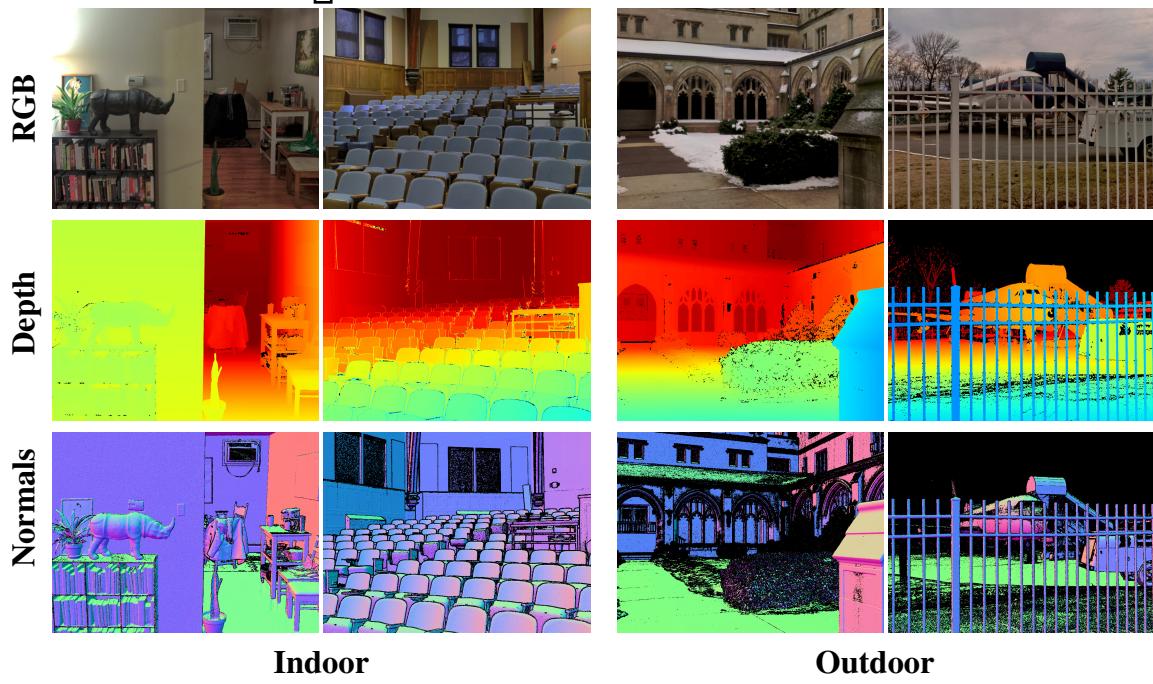


Figure 1: Samples from our DIODE data set. Black represents no valid depth or no valid normals.

Igor Vasiljevic Nick Kolkin Shanyi Zhang[#] Ruotian Luo Haochen Wang[†] Falcon Z. Dai
 Andrea F. Daniele Mohammadreza Mostajabi Steven Basart[†] Matthew R. Walter
 Gregory Shakhnarovich
 TTI-Chicago, [†]University of Chicago, [#]Beihang University

{ivas,nick.kolkin,syzhang,rluo,whc,dai,afdaniele,mostajabi,steven,mwalter,greg}@ttic.edu

Abstract

We introduce DIODE, a dataset that contains thousands of diverse high resolution color images with accurate, dense, long-range depth measurements. DIODE (Dense Indoor/Outdoor DEpth) is the first public dataset to include RGBD images of indoor and outdoor scenes obtained with one sensor suite. This is in contrast to existing datasets that focus on just one domain/scene type and employ different sensors, making generalization across domains difficult.

高清、密集、长程的深度数据，有室内、室外各种场景

1. Introduction

Many of the most dramatic successes of deep learning in computer vision have been for recognition tasks, and relied upon large, diverse, manually labeled datasets such as ImageNet [4], Places [15] and COCO [9]. In contrast, RGBD datasets that pair images and depth cannot be created with crowd-sourced annotation, and instead rely on 3D range sensors that are noisy, sparse, expensive, and often all of the above. Some popular range sensors are specialized for indoor scenes only, due to range limits and sensing technology. Other types of sensors are typically deployed only outdoors. As a result, each of the available RGBD datasets [7, 13, 14, 12] primarily includes only one of these scene types. Furthermore, RGBD datasets tend to be fairly

homogeneous, particularly for outdoor scenes, where the dataset is usually collected with a autonomous driving in mind [7]. While there have been many recent advances in 2.5D and 3D vision, we believe progress has been hindered by the lack of large and diverse real-world datasets comparable to ImageNet and COCO for semantic object recognition.

Depth information is integral to many problems in robotics, including mapping, localization and obstacle avoidance for terrestrial and aerial vehicles, and in computer vision, including augmented and virtual reality [11]. Compared to depth sensors, monocular cameras are inexpensive and ubiquitous, and would provide a compelling alternative if coupled with a predictive model that can accurately estimate depth. Unfortunately, no existing public dataset exists that would allow fitting the parameters of such a model using depth measurements taken by the same sensor in both indoor and outdoor settings. Even if one’s focus is on unsupervised learning of depth perception [8] it is important to have extensive, diverse data set with depth ground truth for evaluation of models.

Indoor RGBD datasets are usually collected using structured light cameras, which provide dense, but noisy, depth maps up to approximately 10m, limiting their application to small indoor environments (e.g., home and office environments). Outdoor datasets are typically collected with a specific application in mind (e.g., self-driving vehicles), and generally acquired with customized sensor arrays consisting of monocular cameras and LiDAR scanners. Typical LiDAR scanners have a high sample rate, but relatively low spatial resolution. Consequently, the characteristics of available indoor and outdoor depth maps are quite different (see Table 1), and networks trained on one kind of data typically generalize poorly to another [6, 10]. Confronting this challenge has attracted recent attention, motivating the CVPR 2018 Robust Vision Challenge workshop.

This paper presents the DIODE (Dense Indoor/Outdoor DEpth) dataset in an effort to address the aforementioned limitations of existing RGBD datasets. DIODE is a large-scale dataset of diverse indoor and outdoor scenes collected using a survey-grade laser scanner (FARO Focus S350 [1]). Figure 1 presents a few representative examples from DIODE, illustrating the diversity of the scenes and the quality of the 3D measurements. This quality allows us to produce not only depth maps of unprecedented density and resolution, but also to derive surface normals with a level of accuracy not possible with existing datasets. The most important feature of DIODE is that it is **the first dataset that covers both indoor and outdoor scenes in the same sensing and imaging setup**.

2. Related Work

A variety of RGBD datasets in which images (RGB) are paired with associated depth maps (D) have been proposed through the years. Most exclusively consist of either indoor or outdoor scenes, and many are tied to a specific task (e.g., residential interior modeling or autonomous driving).

Perhaps the best known RGBD dataset is KITTI [7]. It was collected using a vehicle equipped with a sparse Velodyne VLP-64 LiDAR scanner and RGB cameras, and features street scenes in and around the German city of Karlsruhe. The primary application of KITTI involves perception tasks in the context of self-driving. Thus, the diversity of outdoor scenes is much lower than that of DIODE, but the extent of the street scenes makes it complementary.

Make3D [13] provides RGB and depth information for outdoor scenes that are similar in nature to our dataset, but the depth maps are very low resolution (see Table 1), and there are too few examples for training a deep learning model.

The NYUv2 dataset [14] is widely used for monocular depth estimation in indoor environments. The data was collected with a Kinect RGBD camera, which provides sparse and noisy depth returns. These returns are generally inpainted and smoothed before they can be used for monocular depth estimation tasks. As a result, while the dataset include sufficient samples to train modern machine learning pipelines, the “ground-truth” depth does not necessarily correspond to true scene depth. Our dataset complements NYUv2 by providing very high-resolution, low-noise depth maps of both indoor and outdoor scenes.

Meanwhile, the recent Matterport3D [2] and ScanNet [3] datasets offer a large number of dense depth images of indoor scenes. The datasets were collected from multiple views using a SLAM pipeline. As a result, the depth maps are much noisier and of lower resolution than DIODE, and are intended for semantic tasks like 3D segmentation, rather than accurate 3D reconstruction or depth estimation.

To summarize, compared to existing RGBD datasets, DIODE offers larger scene variety; higher image and depth map resolution; higher density and accuracy of depth measurements; and most importantly, the ability to consider depth perception for indoors and outdoors in a truly unified framework.

3. The DIODE Dataset

We designed and acquired the DIODE dataset with three primary desiderata in mind. First, the dataset should include a diverse set of indoor (e.g., homes, offices, lecture halls, and communal spaces) and outdoor (e.g., city streets, parking lots, parks, forests, and river banks) scenes. Second, the dataset should provide dense depth maps, with accurate short-, mid-, and long-range depth measurements for

a large fraction of image pixels. Third, the depth measurements should be highly accurate.

3.1. Data Acquisition

The aforementioned qualities preclude measuring depth using structured light cameras, and instead requires using a LiDAR. We collected our dataset using a FARO Focus S350 scanner. The FARO is an actuated survey-grade phase-shift laser scanner for both indoor and outdoor environments that provides highly accurate depth measurements over a large depth FOV (between 0.6 m and 350 m with error as low as 1 mm), and at high angular resolution (0.009°). The FARO includes a color camera mounted coaxially with the depth laser, and produces a high-resolution panorama that is automatically aligned with the FARO’s depth returns. These attributes give the FARO a variety of advantages over the more frequently used Velodyne LiDAR with a separate RGB camera, or Kinect depth cameras:

- the scanner is equally well-suited for in indoor and outdoor scanning;
- the point clouds are orders of magnitude more dense;
- the RGB camera is placed very close to the sensor so that there is virtually no baseline between the detector and the camera;

Scanning parameters The FARO allows the customization of various parameters that govern the scanning process. These include the resolution of the resulting depth scan (i.e., the number of points), color resolution of the RGB panorama (i.e., standard or high definition), the quality of the scan (i.e., the integration time of each range measurement). We chose the following scanning settings:

- 1× quality: single scanning pass for every azimuth;
- 360 degree horizontal FOV, 150 degree vertical FOV;
- 1/2 resolution: $\approx 170M$ points;
- 3× HDR: low exposure, regular, high exposure bracketing for RGB.

These settings result in a scan time of approximately 11 min. The intermediate output of a scan is a 20700×8534 (approximately) RGB panorama and a corresponding point cloud, with each 3D point associated with a pixel in the panorama (and thus endowed with color). As with other LiDAR sensors, highly specular objects as well as those that are farther than 350 m (including the sky) do not have an associated depth measurement. Another limitation of the scanner for RGBD data collection is that the LiDAR “sees” through glass or in darkness, resulting in detailed depth maps for image regions that lack the corresponding appearance information.

Scanning Locations We chose scan locations to ensure diversity in the dataset as well a similar number of indoor and outdoor scenes. The scenes include small student offices, large residential buildings, hiking trails, meeting halls, parks, city streets, and parking lots, among others. The scenes were drawn from three different cities. Given the relatively long time required for each scan (approximately 11 min) and the nature of the scanning process, we acquired scans when we could avoid excessive motion and dynamic changes in the scene. However, occasional movement through the scenes is impossible to completely avoid.

The resulting scans exhibit diversity not just between the scenes themselves, but also in the scene composition. Some outdoor scans include a large number of nearby objects (compared to KITTI, where the majority of street scans have few objects near the car), while some indoor scenes include distant objects (e.g., as in the case of large meeting halls and office buildings with large atria), in contrast to scenes in other indoor datasets collected with comparatively short-range sensors.

3.2. Data Curation and Processing

Image Extraction We process the scans to produce a set of rectified RGB images (henceforth referred to as “crops”) at a resolution of 768×1024 . The crops correspond to a grid of viewing directions, at four elevation angles (-20° , -10° , 0° , 10° , 20° , and 30°), and at regular 10° azimuth intervals, yielding 216 viewing directions. We rectify each crop corresponding to 45° (vertical) $\times 60^\circ$ (horizontal) FOV.¹

Curved sections of the panorama corresponding to each viewing frustum must be undistorted to form each rectified crop, i.e., a rectangular image with the correct perspective. To accomplish this we associate each pixel in the rectified crop with a ray (3D vector) in the canonical coordinate frame of the scanner. We use this information to map from panorama pixels and the 3D point cloud to crop pixels.

For each pixel p_{ij} in the desired 768×1024 crop, let the ray passing through the pixel be r_{ij} . We assign the RGB value of each pixel p_{ij} to the average of the RGB values of the nearest five pixels in terms of the angular distance between their rays and r_{ij} .

We employ a similar procedure to generate a rectified depth map. For each ray r_{ij} , we find in the pointcloud the set of 3D points X_{ij} whose rays are nearest to r_{ij} in angular distance.

We discard points with angular distance to r_{ij} greater than 0.5° . We then set the depth of pixel p_{ij} to the robust mean of the depth of points in X_{ij} , using the median 80% of depth values.

¹In the CVPR2019 Workshop version of the paper, we described extracting crops for 67.5° (vertical) $\times 90^\circ$ (horizontal) FOV. That version of the dataset is now deprecated, but available upon request.

	DIODE	NYUv2	KITTI	MAKE3d
Return Density (Empirical)	99.6%/66.9%	68%	16%	0.38%
# Images Indoor/Outdoor	8574/16884	1449/0	0/94000	0/534
Sensor Depth Precision	± 1 mm	± 1 cm	± 2 cm	± 3.5 cm
Sensor Angular Resolution	0.009°	0.09°	0.08°H, 0.4°V	0.25°
Sensor Max Range	350 m	5 m	120 m	80 m
Sensor Min Range	0.6 m	0.5 m	0.9 m	1 m

Table 1: Statistics of DIODE compared to other popular RGBD datasets. Density percentages broken out for indoor and outdoor for DIODE.

In the event that the set X_{ij} is empty we record p_{ij} as having no return (coded as depth 0).

To compute normals for each crop we begin by associating each pointcloud point with a spatial index in the panorama. Then for each spatial index (i, j) of the panorama we take the set of 3d points \hat{X}_{ij} indexed by the 11x11 grid centered on (i, j) , and find a plane using RANSAC [5] which passes through the median of the \hat{X}_{ij} , and for which at least 40% of the points in \hat{X}_{ij} have a residual less than 0.1 cm. We define the normal at position (i, j) to be the vector normal to this plane that faces towards the pointcloud’s origin. Finally for each crop we rotate these normals according to the camera vector, and rectify them via the same procedure used for the depth map.

Crop selection The scanner acquires the full 3D pointcloud before capturing RGB images. This, together with the relatively long scan duration can result in mismatches between certain RGB image regions and the corresponding depth values for dynamic elements of the scene (e.g., when a car present and static during the 3D acquisition moves before the RGB images of its location are acquired). Additionally, some crops might have almost no returns (e.g., an all-sky crop for an outdoor scan). We manually curated the dataset to remove such crops, as well as those dominated by flat, featureless regions (e.g., a bare wall surface close to the scanner).

Masking Though the depth returns are highly accurate and dense, the scanner has some of the same limitations as many LiDAR-based scanners—i.e. erroneous returns on specular objects, “seeing through” glass and darkness causing inconsistencies between RGB and depth, etc.

To ameliorate issues caused by spurious returns, for every crop we create an automated “validity mask” using a robust median filter that rejects depth returns that are too far from the median of a small neighborhood. We provide the raw depth returns to allow users to implement alternative masking or inpainting schemes (e.g. [14]). In addition, for the validation set we manually mask regions with spurious depth or inconsistencies between RGB and depth.

Standard Split We establish a train/validation/test split in order to ensure the reproducibility of our results as well as

to make it easy to track progress of methods using DIODE. The validation set consists of curated crops from 10 indoor and 10 outdoor scans, while the test set consists of crops from 20 indoor and 20 outdoor scans.

When curating scans in the validation and test partitions, we do not allow the fields-of-view of the selected crops to overlap by more than 20° in azimuth for validation scans, and 40° for test scans. No such restriction is used when selecting train crops.

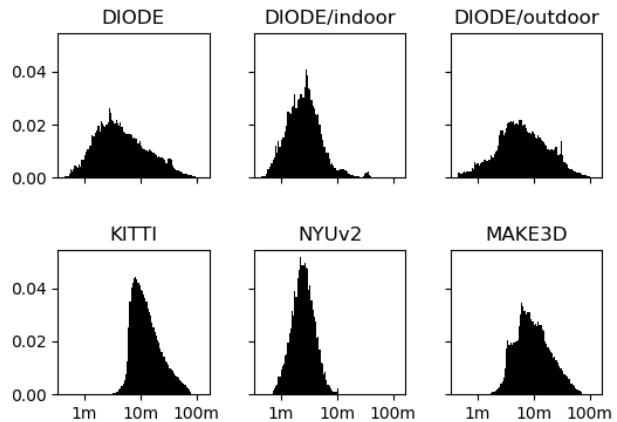


Figure 2: Distribution of measured depth values for DIODE and other popular RGBD datasets.

3.3. Dataset Statistics

Table 1 compares the statistics of DIODE to other widely used RGBD datasets. Note the return density of the data, i.e., the ratio of color pixels with depth measurements to all color pixels; the captured point cloud has a higher resolution than our projected depth maps and thus we have returns for most pixels, missing returns on either very far regions (e.g. sky) or specular regions in indoor images. The depth precision allows for the capture of fine depth edges as well as thin objects.

Figure 2 compares the distribution of values in the depth maps in popular datasets to DIODE (values beyond 100m are only found in DIODE and thus we clip the figures for DIODE for ease of comparison). Note that given that there are often objects both near and far from the camera in outdoor scans, the distribution of depth values is more diffuse in DIODE/outdoor than in KITTI. Only the much smaller

and lower resolution Make3D is close to matching the diversity of DIODE depth values.

4. Conclusion

We expect the unique characteristics of DIODE, in particular the density and accuracy of depth data and above all the unified framework for indoor and outdoor scenes, to enable more realistic evaluation of depth prediction methods and facilitate progress towards general depth estimation methods. We plan to continue acquiring additional data to expand DIODE, including more locations and additional variety in weather and season.

References

- [1] FAROT™ S350 scanner. <https://www.faro.com/products/construction-bim-cim/faro-focus/>. Accessed: 2019-03-20.
- [2] 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. *arXiv preprint arXiv:1709.06158*, 2017.
- [3] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.
- [4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [5] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.
- [6] R. Garg, V. K. B.G., G. Carneiro, and I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.
- [7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 2013.
- [8] V. Guizilini, R. Ambrus, S. Pillai, and A. Gaidon. Packnet-sfm: 3d packing for self-supervised monocular depth estimation. *arXiv preprint arXiv:1905.02693*, 2019.
- [9] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.
- [10] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE PAMI*, 2016.
- [11] E. Marchand, H. Uchiyama, and F. Spindler. Pose estimation for augmented reality: a hands-on survey. *TVCG*, 2016.
- [12] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.
- [13] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Depth perception from a single still image. In *AAAI*, 2008.
- [14] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.
- [15] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014.