Joint 2D-3D-Semantic Data for Indoor Scene Understanding

Iro Armeni[1*] Alexander Sax[1*] Amir R. Zamir[1,2] Silvio Savarese[1]
[1] Stanford University [2] University of California, Berkeley

http://3Dsemantics.stanford.edu/

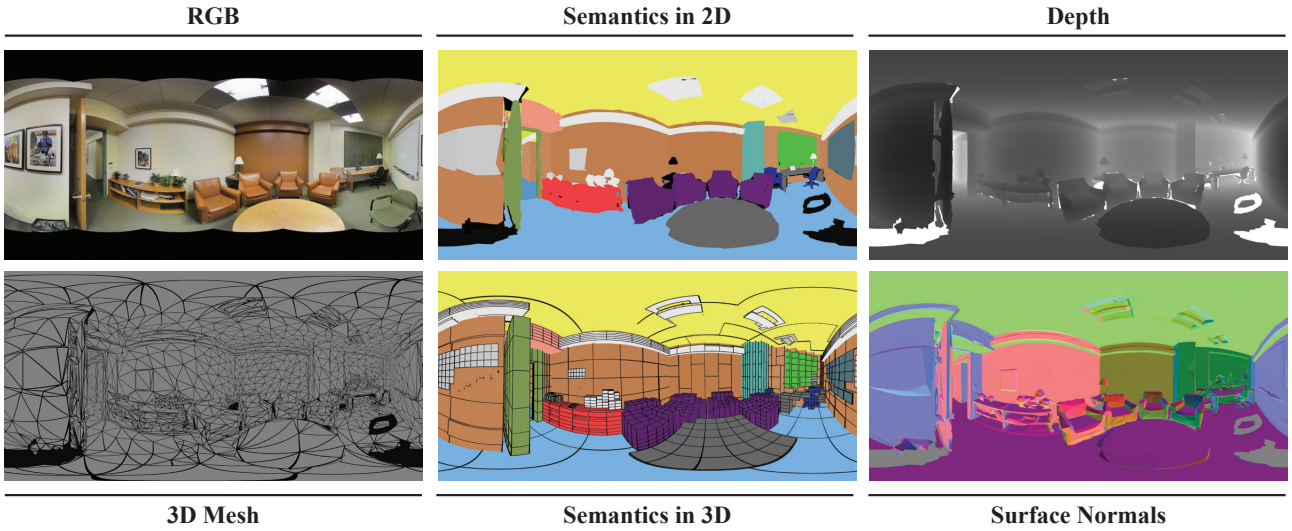

Figure 1: **Joint 2D-3D-Semantic data.** We present a dataset that provides a variety of mutually registered modalities including: RGB images, depth, surface normals, global XYZ images as well as instance-level semantic annotations in 2D and 3D. The modalities are in 2D (*e.g.* RGB images), 2.5D (*e.g.* depth) and 3D (*e.g.* meshes).

Abstract

We present a dataset of large-scale indoor spaces that provides a variety of mutually registered modalities from 2D, 2.5D and 3D domains, with instance-level semantic and geometric annotations. The dataset covers over 6,000 m^2 and contains over 70,000 RGB images, along with the corresponding depths, surface normals, semantic annotations, global XYZ images (all in forms of both regular and 360° equirectangular images) as well as camera information. It also includes registered raw and semantically annotated 3D meshes and point clouds. The dataset enables development of joint and cross-modal learning models and potentially unsupervised approaches utilizing the regularities present in large-scale indoor spaces.

1. Introduction

There is an important advantage in analyzing 3D data, especially ones that originate from comprehensive large-scale scans: the entire geometry of an object and its surrounding context are available at once. This can provide strong cues for semantics, layout, occlusion handling, shape completion, amodal detection, etc. This rich geometric information is complementary to the RGB domain, which offers dense appearance features. Hence, there is a great potential in developing models that perform joint or cross-modal learning to enhance the performance. Although several RGB-D [6, 7, 16, 15] and a few 3D datasets [12, 13, 14, 1] have been developed to date, the majority of them are limited in the scale, diversity, and/or the number of modalities they provide.

Over the past few years, advances in the field of 3D imaging have led to manufacturing inexpensive sensors and mainstreaming their use in consumer products (*e.g.* Kinect [5], Structure Sensor [3], RealSense [4], etc). The Computer Vision community has been affected by this change and is experiencing a lot of development in data-driven 2.5D and 3D vision [6, 17, 18, 19, 1]. However, the 3D sensing field has recently undergone a follow-up shift with the availability of mature technology for scan-

* Both authors contributed equally.

Table 1: Comparison of existing 2.5D and 3D Datasets.

Dataset	**Stanford Scenes** [12]	**SceneNet** [13]	**SceneNet RGBD** [15]	**SUNCG** [14]	**NYUD2** [6]	**SUN RGBD** [7]	**SceneNN** [16]	**2D-3D-S (Ours)**
Type of Data	Synthetic	Synthetic	Synthetic	Synthetic	Real	Real	Real	Real
RGB	-	-	5M	-	1,449	10,335	-	70,496
Depth	✗	✗	✓	130,269	✓	✓	✓	✓
Collection Method	✗	✗	Rendered Video	Rendered Depth Images	Video	Video	Video	360° scan
Surf. Normals	✗	✗	✗	✗	✓	✓	✓	✓
2D Semantics	✗	✗	✓	✗	✓	✓	✓	✓
Resolution	-	-	320×240	-	640×480	640×480	640×480	1080×1080
3D Point Cloud (PC)	✗	✗	✗	✗	✓	✓	✗	✓
3D Mesh /CAD	✓	✓	✗	✓	✗	✗	✓	✓
3D Semantic Mesh/ CAD	✓	✓	✗	✓	✗	✗	✓	✓
# Object Class	-	-	255	84	894	800	-	13
# Scene Categories	-	5	5	24	26	47	-	11
# Scene Layouts	130	57	57*	45,622	464	-	100	270

✗: not included, ✓: included, -: information not available, *:16,895 configurations

ning *large-scale* spaces, *e.g*. an entire building. Such systems can reliably form the 3D point cloud of thousands of square meters with the number of points often exceeding hundreds of millions. This demands developing methods capable of coping with this scale, and ideally, exploiting the unique characteristics of such data.

To enable parsing the aforementioned goals, we present a 2D-3D semantic dataset that can be used for a plethora of tasks, such as scene understanding, depth estimation, surface normals estimation, object detection, segmentation, amodal detection, and scene reconstruction. Also, the 3D mesh models and equirectangular projections can be used to generate a virtually unlimited number of images, something that is currently possible only in synthetic 3D datasets. The dataset along with 360° visualizations is available for download at http://3Dsemantics.stanford.edu/.

2. Related Datasets

There exist several RGB-D datasets in the literature related to scene understanding; NYU Depth v2[6], SUN RGBD[7] and recently SceneNN [16] are among the prominent ones. The latter provides a larger number of images than the rest, though the increased number originates from densely annotated frames of videos from a smaller number of scenes. Due to the complexity of collecting and densely annotating such data, many synthetic datasets have appeared lately (RGBD: [15], 3D: [12, 13, 14]). Their advantage is that by employing large-scale object libraries (*e.g*. ShapeNet [11]) one could generate a virtually infinite number of images along with the corresponding semantics.

Existing real datasets are relatively limited to one particular task and the 2.5D domain. Recently, [16] offered watertight mesh models of the reconstructed scenes and [1] presented a 3D point cloud dataset of large-scale indoor spaces. Although such data can be used for more than semantic detection/segmentation, it is not suitable for tasks across different data dimensionalities and modalities.

The proposed 2D-3D dataset includes RGB, depth, equirectangular and global XYZ OpenEXR images, as well as 3D meshes and point clouds of the same indoor spaces (Table 1). The different modalities can be used independently or jointly to develop learning models that seamlessly transcend across domains. Also, using the provided equirectangular images, camera parameters, and 3D mesh models, it is possible to generate additional data tailored to specific tasks. In comparison to existing real-world datasets, it offers a greater number of images as well as additional modalities, such that of equirectangular images or 3D meshes. It also provides consistent annotations across all modalities and dimensions. However, it is currently limited in the number of object and scene categories.

3. Dataset Overview

The dataset is collected in 6 large-scale indoor areas that originate from 3 different buildings of mainly educational and office use. For each area, all modalities are registered in the same reference system, yielding pixel to pixel correspondences among them. In a nutshell, the presented dataset contains a total of 70,496 regular RGB and 1,413 equirectangular RGB images, along with their corresponding depths, surface normals, semantic annotations, global XYZ OpenEXR format and camera metadata. In addition, we provide whole building 3D reconstructions as textured meshes, as well as the corresponding 3D semantic meshes. We also include the colored 3D point cloud data of these areas with the total number of 695,878,620 points, that has been previously presented in the Stanford large-scale 3D Indoor Spaces Dataset (S3DIS [1]). The annotations are instance-level, and consistent across all modalities and correspond to 13 object classes. We refer the readers to Tables 7 and 3 for statistics on the scene and object categories. Figure 3 shows an example of the equirectangular images for one scan.

4. Collection and Processing

We collected the data using the Matterport Camera [2], which combines 3 structured-light sensors to capture 18 RGB and depth images during a 360° rotation at each

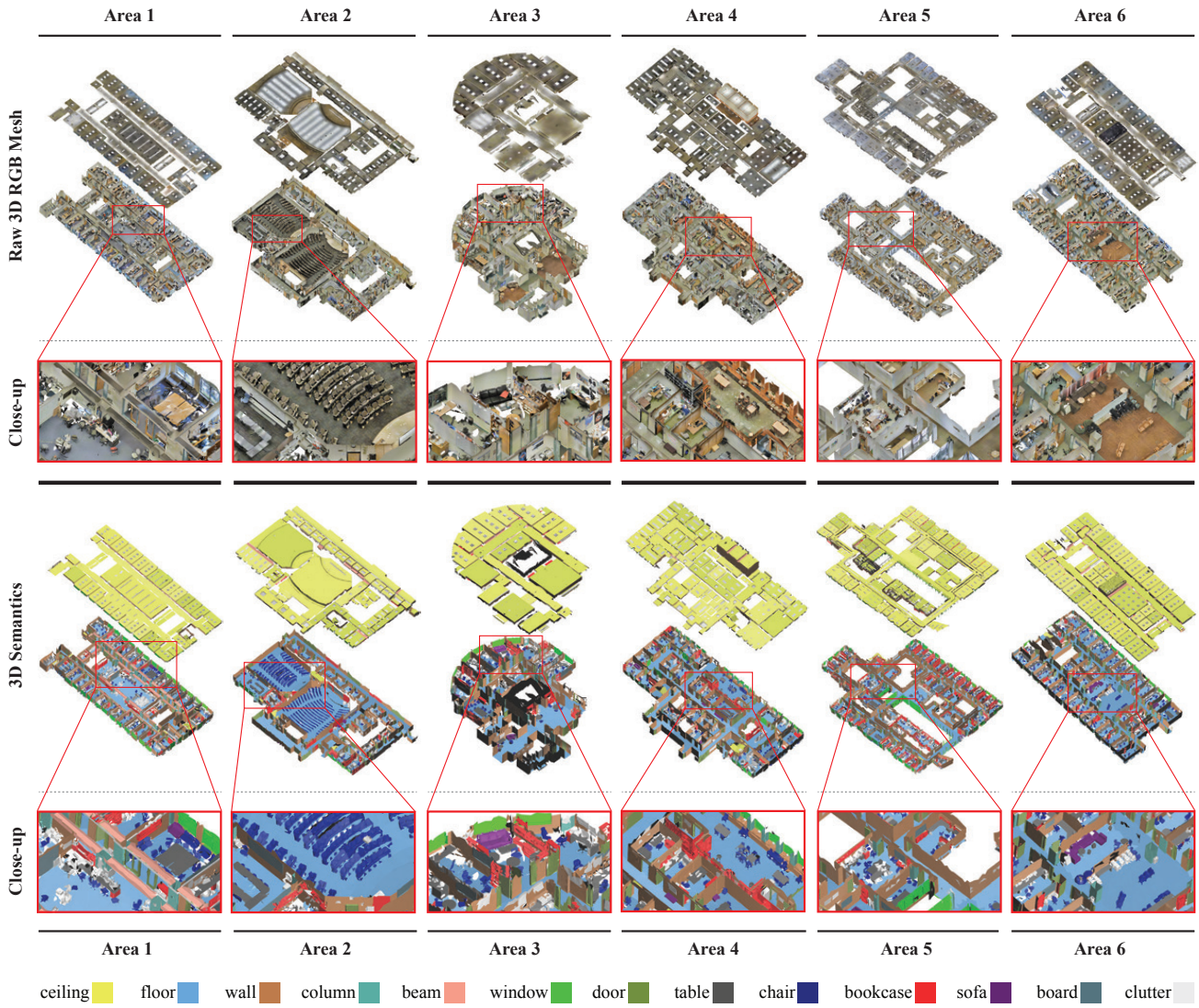

Figure 2: **3D Modalities.** The dataset includes both the textured and semantic 3D mesh models of all areas as well as their point clouds.

scan location. The output is the reconstructed 3D textured meshes of the scanned area, the raw RGB-D images, and camera metadata. We used this data as a basis to generate additional RGB-D data and make point clouds by sampling the meshes. We semantically annotated the data directly on the 3D point cloud, rather than images, and then projected the per point labels on the 3D mesh and the image domains.

The rest of the section elaborates on each modality.

4.1. 3D modalities

The dataset contains two main 3D modalities (3D point cloud data and 3D mesh model) and their semantic counterparts for each of the 6 areas. Statistics related to this modality are offered in Table 2.

3D Point Cloud and Mesh: As mentioned above, we receive the reconstructed 3D textured Mesh model for each scanned area from the Matterport Camera. Each model contains an average of 200k triangulated faces and a material mapping to texture images providing a realistic reconstruction of the scanned space. We generate the colored 3D point clouds by densely and uniformly sampling points on the mesh surface and assigning the corresponding color.

3D Semantics (Labeled Mesh and Voxels): We semantically annotate the data on the *3D point cloud* and assign one of the following 13 object classes on a per-point basis: *ceiling, floor, wall, beam, column, window, door, table, chair, sofa, bookcase, board* and *clutter* for all other elements. Performing annotations in 3D, rather than 2D, provides 3D object models and enables performing occlusion and amodal analysis, yet the semantics can be projected

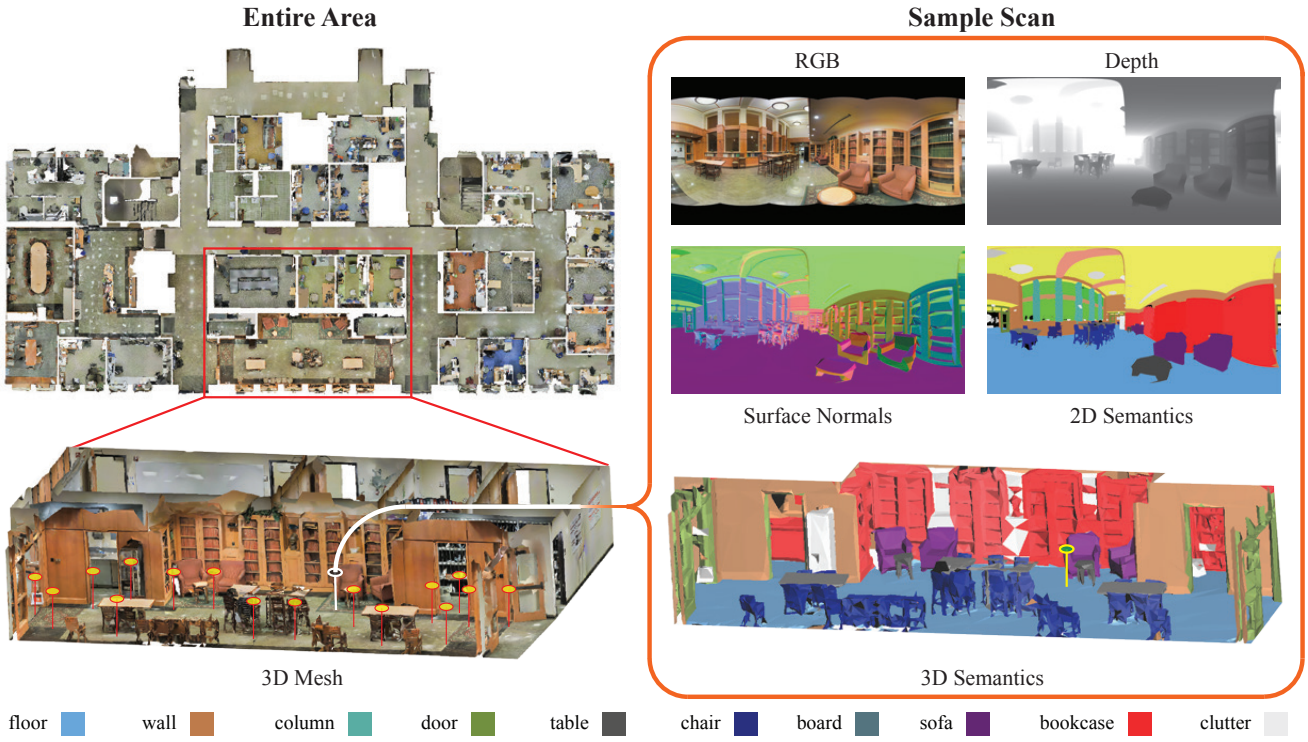

Figure 3: **Data Processing.** The RGB output of the scanner is registered on the 3D modalities (each yellow marker represents one scan). We then process the RGB and 3D data to make depth, surface normals, and 2D semantic (projected from 3D semantics) images for each scan. The processed equirectangular images are sampled to make new regular images (shown in Figure 4).

Table 2: Statistics of 3D Data

Area	Number of 3D Points	Number of Mesh Faces
1	43,956,907	158,500
2	470,023,210	361,830
3	18,662,173	147,420
4	43,278,148	201,735
5	78,649,818	198,220
6	41,308,364	198,590
Total	**695,878,620**	**1,266,295**

onto any number of images to provide ground truth annotations in 2D as well. Each object instance in the dataset has a unique identifier. We also annotate the point cloud data into rooms and assign one of the following 11 scene labels to each: *office, conference room, hallway, auditorium, open space, lobby, lounge, pantry, copy room, storage* and *WC*. Again, each instance in the point cloud receives a unique index. Given these annotations, we calculate the tightest axis-aligned object bounding box of each instance and further voxelize it into a $6 \times 6 \times 6$ grid with binary occupancy. This information provides a better understanding of the underlying geometry and can be leveraged, for example, in 3D object detection or classification.

We then project the object and scene semantics on the mesh model's faces and generate 3D semantic meshes that preserves the same class structure and instance index. We used a voting scheme to transfer these annotations to the mesh. Each annotated point casts a vote for the face that is nearest to it, then votes are tallied and each face is annotated with the mode class. Faces which garner no votes belong to non-annotated parts of the dataset and are labeled as the default $<UNK>\_0\_<UNK>\_0\_0$ class (null). Our 3D models are labeled with the class of the object and the specific instance. These instances are globally unique among all models and are indexed in $semantic\_labels.json$. Each object is stored in this file as $class\_instanceNum\_roomType\_roomNum\_areaNum$. Note that instances, rooms and areas are 1-indexed so that the singleton $<UNK>$ class is unique in that it has $instanceNum = 0$, $roomNum = 0$ and $areaNum = 0$. Figure 2 shows the raw and semantically annotated 3D mesh models for all 6 areas.

4.2. 2D modalities

The dataset contains densely sampled RGB images per scan location. These images were sampled from equirectangular images that were generated per scan location and modality using the raw data captured by the scanner (also

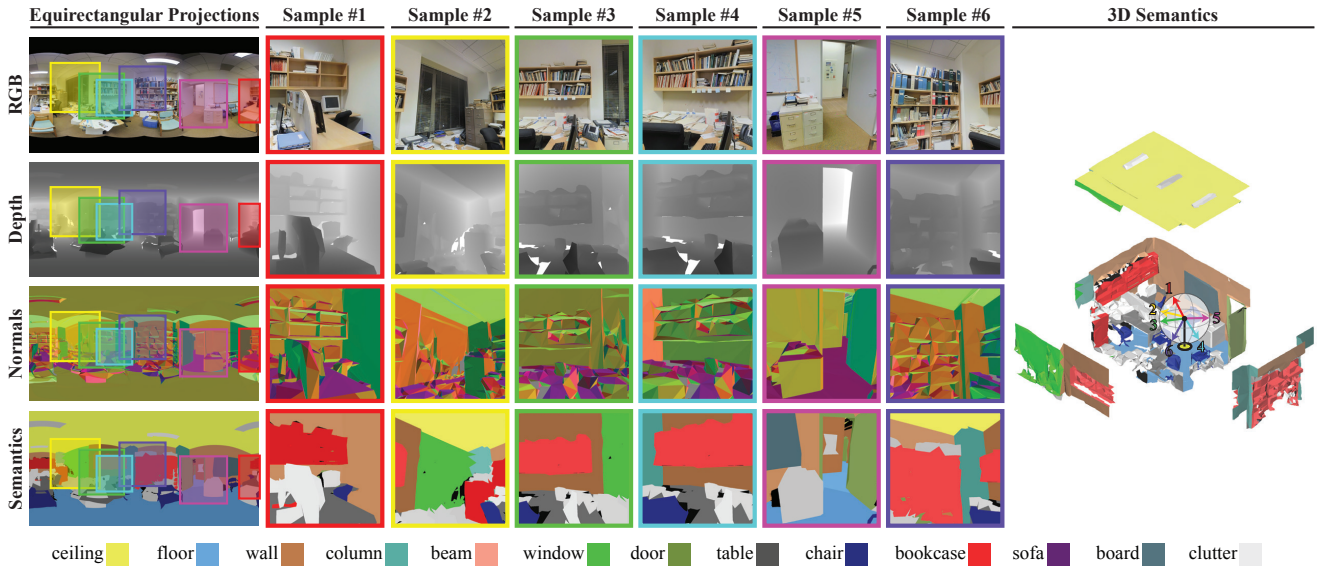

Figure 4: **Sampling images from the equirectangular projection.** We use the equirectangular projections to sample 72 images per scan location, all with consistent depth, surface normal, and semantic information. The sampling distributions are provided in Figure 5.

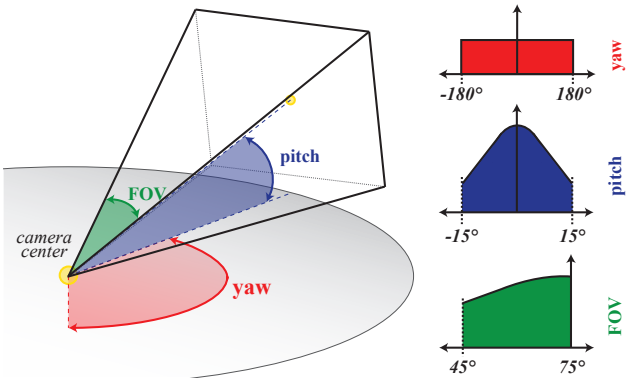

Figure 5: **Sampling distributions.** We sample camera parameters from the above distributions of yaw, pitch and Field of View (FOV).

part of the dataset). Statistics of the 2D modalities are offered in Table 4. All images in the dataset are stored in full high-definition at 1080×1080 resolution.

RGB Images: We use the provided raw RGB data to form a cubemap per scan location and sample new images in this space. We randomly sample the camera's euler angles as follows (Figure 5): (a) yaw: uniform in $[-180°, 180°]$, (b) pitch: Gaussian with zero mean and $15°$ standard deviation, and (c) roll: always zero. Field of View (FOV) angles are sampled from a half Gaussian distribution with $75°$ mean and $-30°$ standard deviation.

To avoid having images that are uninteresting in terms of semantic content (*e.g.* only a plain wall), we generate 3×24 images per scan location and preserve approximately 70% of them by sampling based on semantic content entropy. We perform the entropy sampling as follows: we compute Shannon's entropy for each image on the pixel distribution of semantic classes therein (i.e. a 13 bin distribution). We then discard the bottom $\sim$15% (as such cases correspond to images with very small semantic value, *e.g.* a close up of a wall) and preserve the top $\sim$60% (the semantically diverse images). Out of the rest of the images, we preserve about 50% of them by sampling the entropy values on a half Gaussian with mean and standard deviation of 1 and $-\frac{1}{2}$, respectively.

Using this approach, rather than simply discarding the low-entropy images, preserves the dataset's diversity by not completely removing low-entropy scenes. We rendered all images via Blender 2.78. Figure 4 shows examples of sampled images per modality on an equirectangular image.

Meta-data and Camera Parameters per Image: For each generated image we provide the camera pose in the 'pose folder'.

Depth Images: For each image, we provide the depth, which was computed from the 3D mesh instead of directly from the scanner. We rendered these images from the 3D mesh by saving out depth information from the z-buffer in Blender. The images are saved as 16-bit grayscale PNGs where one unit of change in pixel intensity (e.g. a value from 45 to 46) corresponds to a $\frac{1}{512}m$ change in depth. The maximum observable range is therefore about 128 meters ($2^{16} \cdot \frac{1}{512}$). All depths beyond this maximum distance assume the maximum value (65,535). Pixels that correspond to locations where there is no depth information also take this maximum distance.

Surface Normal Images: The surface normals were

Table 3: Object Class Statistics

	Structural Elements							Movable Elements					Total
Area	ceiling	floor	wall	beam	column	window	door	chair	table	bookcase	sofa	board	
1	56	45	235	62	58	30	87	156	70	91	7	28	**925**
2	82	51	284	12	20	9	94	546	47	49	7	18	**1,219**
3	38	24	160	14	13	9	38	68	31	42	10	13	**460**
4	74	51	281	4	39	41	108	160	80	99	15	11	**963**
5	77	69	344	4	75	53	128	259	155	218	12	43	**1,437**
6	64	50	248	69	55	32	94	180	78	91	10	30	**1,001**
Total	**391**	**290**	**1,552**	**165**	**260**	**174**	**549**	**1,369**	**461**	**590**	**61**	**143**	**6,005**

Table 4: Statistics of Images

Area	# of Images per 2/2.5D Modality		Total
	Image Type **I**	Image Type **E**	
1	10,327	190	42,068
2	15,714	299	64,052
3	3,704	85	15,156
4	13,268	258	54,104
5	17,593	373	71,864
6	9,890	208	40,392
Total	70,496	1,413	287,636

I: Regular Images, **E**: Equirectangular Images

Table 5: **Training and Testing Splits (3-fold cross-validation)**

Fold #	Training (Area #)	Testing (Area #)
1	1, 2, 3, 4, 6	5
2	1, 3, 5, 6	2, 4
3	2, 4, 5	1, 3, 6

computed from a normals pass in Blender and are saved as 24-bit RBG PNGs. The surface normals in 3D corresponding to each pixel are computed from the 3D mesh instead of directly from the depth image. The normal vector is saved in the RGB color value where Red is the horizontal value (more red to the right), Green is vertical (more green downwards), and Blue is towards the camera. Each channel is 127.5-centered, so both values to the left and right (of the axis) are possible. For example, a surface normal pointing straight at the camera would be colored (128, 128, 255) since pixels must be integer-valued.

Missing values take (128,128,128) which is convenient in practice as it is not a unit normal and is clearly visually distinguishable from the surrounding values. The convention is that surface normals cannot point away from the camera.

Semantically Labeled Images: We project the 3D semantics from the mesh model onto the 2D image. Due to certain geometric artifacts present at the mesh model mainly because of the level of detail in the reconstruction, the 2D annotations occasionally present small local misalignment to the underlying pixels, especially for points that have a short distance to the camera. This issue can be easily addressed by fusing image content with the projected annotations using graphical models.

The semantically labeled images are saved as 24-bit RGB PNGs, but each pixel's color value can be directly interpreted as an index into the list. For example, $board\_3\_hallway\_4\_6$ is at index 257 in $semantics\_labels.json$. Since 257 equals #000101 in hex, #000101 is the color of the chair in the image. For semantic images, pixel values that correspond to holes in the mesh contain the value (13, 13, 13).

3D Coordinate Encoded Images: The pixels in these images encode the X, Y, Z location of the point in the world coordinate system. This information can be used for conveniently relating the content of the RGB images, *e.g.* forming correspondences. The images are stored in the OpenEXR format with each channel contiaining 16-bit floating point numbers. Utility functions, including for reading EXRs, are provided in the GitHub repo accessible on the website.

4.3. Naming Convention

```
README.md
/assets
  semantic_labels.json
  utils.py
/area_1
  /3d
    pointcloud.mat
    rgb.obj      # The raw 3d mesh with rgb textures
    rgb.mtl      # The textures for the raw 3d mesh
    semantic.obj # Semantically-tagged 3d mesh
    semantic.mtl # Tectures for semantic.obj
    /rgb_textures
      {uuid_{i}}.jpg  # Texture images for the rgb 3d mesh
  /data   # all of the generated data
    /pose
      camera_{uuid}__{room}_{i}_frame_{j}_domain__pose.json
    /rgb
      camera_{uuid}__{room}_{i}_frame_{j}_domain__rgb.png
    /depth
    /global_xyz
    /normal
    /semantic
    /semantic_pretty
  /pano    # equirectangular projections
    /pose
      camera_{uuid}__{room}_{i}_frame_equirectangular_domain__pose.json
    /rgb
    /depth
    /global_xyz
    /normal
    /semantic
    /semantic_pretty
/area_2
/area_3
/area_4
/area_5a
/area_5b
/area_6
```

The filenames of images in the dataset are globally unique as no two files share a camera uuid, frame number, and domain. The room type is included for convenient filtering.

Table 6: **Baseline 3D Object Detection Results ([1]).** Class specific average precision (AP) using different features.

	Structural Elements								Movable Elements						overall
	ceiling	floor	wall	beam	column	window	door	mean	table	chair	sofa	bookcase	board	mean	mean
Full model	**71.61**	**88.70**	**72.86**	66.67	**91.77**	**25.92**	**54.11**	**67.38**	46.02	**16.15**	**6.78**	54.71	3.91	**25.51**	**49.93**
No global	48.93	83.76	65.25	62.86	83.15	22.55	41.08	57.27	37.57	11.80	4.57	45.49	3.06	20.35	41.87
No local	50.74	80.48	65.59	**68.53**	85.08	21.17	45.39	58.73	39.87	11.43	4.91	**57.76**	3.73	23.78	44.19
No color	48.05	80.95	67.78	68.02	87.41	25.32	44.31	59.73	**50.56**	11.83	6.32	52.33	**4.76**	25.30	45.41

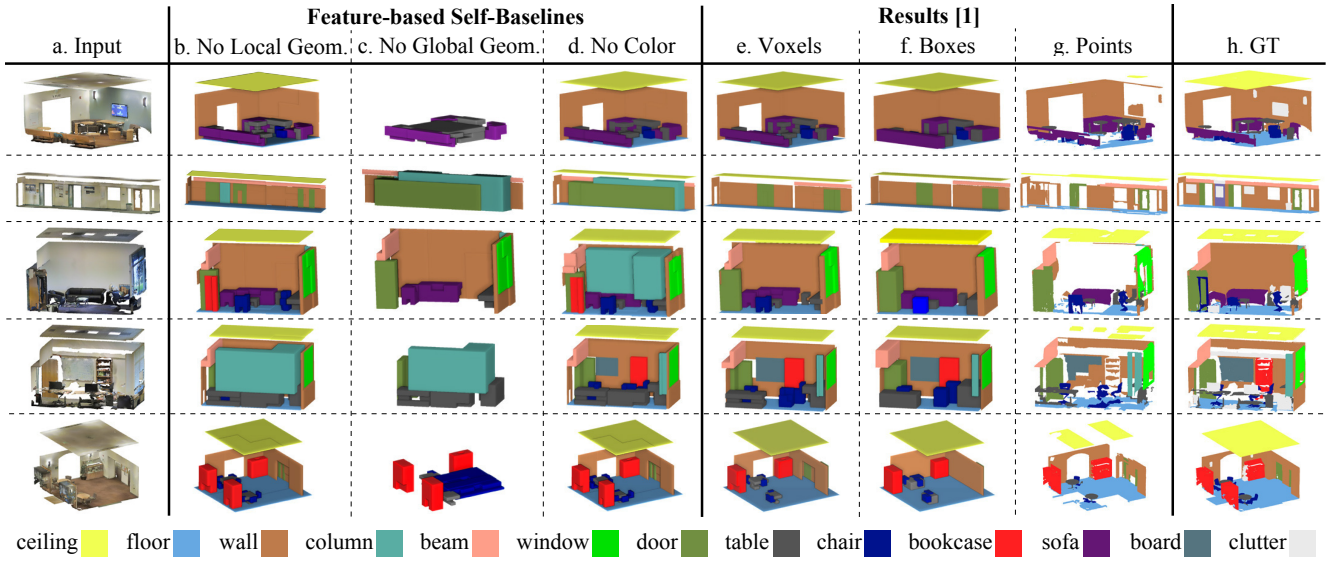

Figure 6: **Qualitative Baseline 3D Object Detection Results ([1]).**

5. Sample Data

Figure 7 provides a representative sample of the generated data showing the diversity of the indoor scenes, in terms of scene category, appearance, intra-class variation, density of objects and amount of clutter. This also shows the varying degree of difficulty of the data, which consists of both easy and hard examples.

6. Train and Test splits

Certain areas in the dataset represent parts of buildings with similarities in their appearance and architectural features, thus we define standard training and testing splits so that no areas from similarly looking buildings appear in both. We split the 6 areas in the dataset as per Table 5 and follow a 3-fold cross-validation scheme.

7. Baseline Results

As a baseline, we provide results on the task of 3D object detection, performed on the 3D point clouds from this paper [1]. The method follows a hierarchical approach to semantic parsing of large-scale data: first, we parse the raw data into semantically meaningful spaces (*e.g.* rooms, etc) and align them into a canonical reference coordinate system. Second, we parse each of these spaces into comprising elements that belong to one of the 12 available classes. We implement the first step with an unsupervised approach, and the second by training one-vs-all SVMs for each object class. We also employ a CRF for contextual consistency. Our experimental setup follows Table 5. For more details we refer the reader to [1]. Table 6 tabulates the quantitative results of this baseline on the proposed dataset. Figure 6 showcases sample qualitative results.

8. Conclusion

We presented a dataset of large-scale indoor spaces. The main property of the dataset is being comprised of mutually registered modalities including RGB images, surface normals, depths, global XYZ images, scene labels, and 2D semantics as well as raw and semantically annotated 3D meshes. The semantic annotations were performed in 3D and were consistently projected across all modalities and dimensions. We provided the 2D and 2.5D modalities in forms of both regular and 360° equirectangular images. Finally, we described our collection, processing, and sampling procedures along with baseline results on 3D ob-

Table 7: Scene and Space Statistics

Area	Sq. meters	Number of Instances Per Scene Category											Total
		Office	Conference Room	Auditorium	Lobby	Lounge	Hallway	Copy Room	Pantry	Open Space	Storage	WC	
1	965	31	2	-	-	-	8	1	1	-	-	1	**45**
2	1,100	14	1	2	-	-	12	-	-	-	9	2	**39**
3	450	10	1	-	-	2	6	-	-	-	2	2	**24**
4	870	22	3	-	2	-	14	-	-	-	4	2	**49**
5	1,700	42	3	-	1	-	15	-	1	-	4	2	**55**
6	935	37	1	-	-	1	6	1	1	1	-	-	**53**
Total	**6,020**	**156**	**11**	**2**	**3**	**3**	**61**	**2**	**3**	**1**	**19**	**9**	**270**

ject detection. We hope the dataset fuels development of cross and joint modality techniques as well as unsupervised approaches leveraging the regularities in large-scale man-made spaces.

References

[1] Iro Armeni and Ozan Sener and Amir R. Zamir and Helen Jiang and Ioannis Brilakis and Martin Fischer and Silvio Savarese. 3D Semantic Parsing of Large-Scale Indoor Spaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[2] Matterport: 3D models of interior spaces. http://matterport.com/. Accessed: 2017-02-02.

[3] Structure Sensor. https://structure.io/. Accessed: 2017-02-02.

[4] Intel RealSense. http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html. Accessed: 2017-02-02

[5] Kinect Sensor. https://developer.microsoft.com/en-us/windows/kinect/develop. Accessed: 2017-02-02.

[6] Nathan Silberman, Derek Hoiem, Pushmeet Kohli and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *Proceedings of the European Conference on Computer Vision (ECCV)* 2012.

[7] Song, Shuran and Lichtenberg, Samuel P and Xiao, Jianxiong. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages=567-576, 2015.

[8] Xiao, Jianxiong and Owens, Andrew and Torralba, Antonio. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages=1625-1632, 2013.

[9] Janoch, Allison and Karayev, Sergey and Jia, Yangqing and Barron, Jonathan T and Fritz, Mario and Saenko, Kate and Darrell, Trevor. A category-level 3d object dataset: Putting the kinect to work. *Consumer Depth Cameras for Computer Vision*, pages=141-165, 2013, Springer.

[10] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages=248-255, 2009.

[11] Chang, Angel X and Funkhouser, Thomas and Guibas, Leonidas and Hanrahan, Pat and Huang, Qixing and Li, Zimo and Savarese, Silvio and Savva, Manolis and Song, Shuran and Su, Hao and others Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015

[12] Fisher, Matthew and Ritchie, Daniel and Savva, Manolis and Funkhouser, Thomas and Hanrahan, Pat. Example-based synthesis of 3D object arrangements. *ACM Transactions on Graphics (TOG)*, Vol. 31, No 6, pages 135, 2012, ACM.

[13] Handa, Ankur and Patraucean, Viorica and Badrinarayanan, Vijay and Stent, Simon and Cipolla, Roberto. Understanding Real World Indoor Scenes With Synthetic Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2016.

[14] Song, Shuran and Yu, Fisher and Zeng, Andy and Chang, Angel X and Savva, Manolis and Funkhouser, Thomas. mantic Scene Completion from a Single Depth Image. Semantic Scene Completion from a Single Depth Image. *arXiv preprint arXiv:1611.08974*, 2016

[15] McCormac, John and Handa, Ankur and Leutenegger, Stefan and Davison, Andrew J. SceneNet RGB-D: 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth. In *arXiv preprint arXiv:1612.05079*, 2016.

[16] Hua, Binh-Son and Pham, Quang-Hieu and Nguyen, Duc Thanh and Tran, Minh-Khoi and Yu, Lap-Fai and Yeung, Sai-Kit. Scenenn: A scene meshes dataset with annotations. In *Proceedings of the Fourth International Conference 3D Vision (3DV)*, pages=92-101, 2016.

[17] Ren, Xiaofeng and Bo, Liefeng and Fox, Dieter. Rgb-(d) scene labeling: Features and algorithms. In *Proceedingd of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages=2759-2766, 2012.

[18] Bo, Liefeng and Lai, Kevin and Ren, Xiaofeng and Fox, Dieter Object recognition with hierarchical kernel descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages=1729-1736, 2011.

[19] Koppula, Hema S and Anand, Abhishek and Joachims, Thorsten and Saxena, Ashutosh Semantic labeling of 3d point clouds for indoor scenes. In *Advances in neural information processing systems*, pages=244-252, 2011

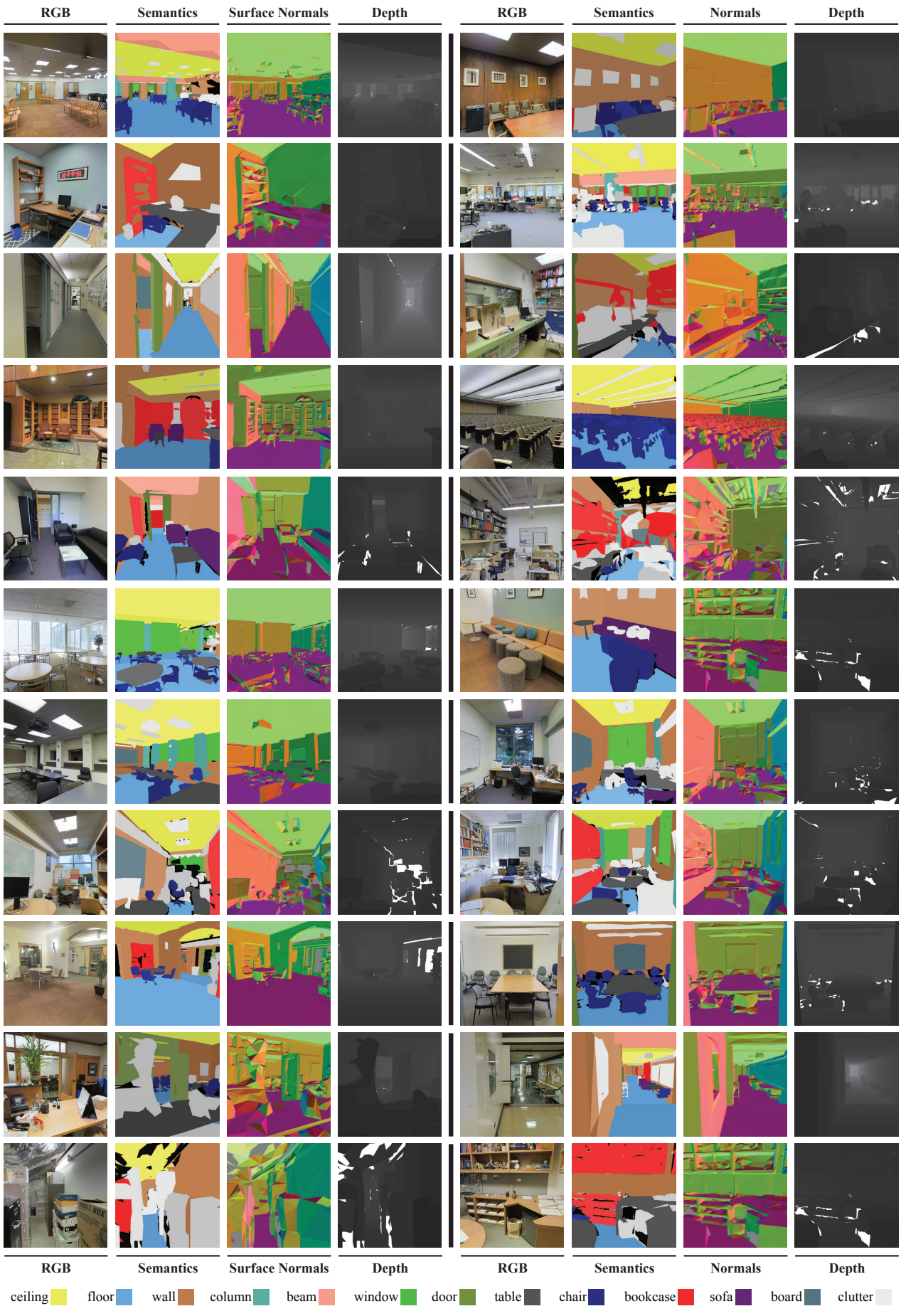

Figure 7: **Examples Images of 2D and 2.5D Modalities.** RGB, Semantic, Surface Normals and Depth images.