ToF-360 – A Panoramic Time-of-flight RGB-D Dataset for Single Capture Indoor Semantic 3D Reconstruction

Hideaki Kanayama [⋆1,2], Mahdi Chamseddine [⋆3,4], Suresh Guttikonda [⋆3],
So Okumura [1,2], Soichiro Yokota [1,2], Didier Stricker [3,4], Jason Rambach [3]

Abstract

3D scene understanding is a key research topic for various automation areas. Many RGB-D datasets today focus on reconstruction of entire scenes. However, their scanning processes are time-consuming, requiring multiple or continuous recordings using a scanner with a limited angle of view. Such datasets often contain data affected by stitching artifacts or poor quality annotation masks projected directly from 3D to image. In this paper, we present ToF-360. This is the first RGB-D dataset obtained by a unique Time-of-Flight (ToF) sensor capable of 360° *omnidirectional RGB-D scanning within seconds. In addition to the raw data in a fisheye format and equi-rectangular projection (ERP) images from the device, we provide manually labeled high-quality, pixel-level, 2D semantics and room layout annotations and introduce a benchmark for three practical tasks: 2D semantic segmentation, 3D semantic segmentation, and layout estimation. We demonstrate that our dataset helps to better represent real-world scenarios and push the limits of existing state-of-the-art methods. The dataset is publicly available at* https://doi.org/10.57967/hf/5074.

1. Introduction

In recent years, there has been increased interest in indoor 3D scene understanding for many practical applications in the domains of augmented- and virtual reality (AR/VR), autonomous driving, scene modeling, and robot navigation [8, 32]. Many recent machine learning techniques including depth estimation, 3D reconstruction, and semantic segmentation [1, 33, 53] have tackled different parts of this challenge. Most of these tasks have progressed with the increased availability of affordable commercial 3D sensing devices, which enabled a variety of RGB-D datasets. However, since datasets such as SUN RGB-D [40], ScanNet [12], and Matterport3D [6] are based on scanning with specific stationary devices, it is still not trivial to make use of them for practical applications. Meanwhile, these devices typically need from tens of seconds up to minutes for capturing and require a static environment to be maintained during scanning. Moreover, some of these datasets exhibit image alignment artifacts or low-quality segmentation mask labels directly projected from 3D (see Figure 1).

Even though mobile scanning LiDAR devices like Microsoft Kinect [10], iPhone LiDAR [22], and the RealSense LiDAR camera [11] facilitate data collection, continuous scanning is required because of the restriction on the angle of the view angle to $< 120^\circ$, which extends the acquisition time. These limitations while recording data can be a barrier to adoption to practical applications. Ricoh released a novel handheld Time-of-Flight (ToF) device capable of capturing complete colored 3D point clouds omnidirectionally from a single camera shot in one second [31]. It solves the problem of portability on conventional stationary devices and restriction of the angle of view on commercial mobile scanners, which leads to shorter and less cumbersome scanning procedures. We expect this scanner to stimulate the development of novel algorithms for single-shot reconstruction tasks that do not require global position alignment and to bridge the gap between research and actual applications.

Our dataset, **ToF-360**, consists of 207 spherical RGB-D images taken in 4 unique environments. We emphasize the precise annotation and superior data quality of our dataset, compared to other datasets and 3D scanners in Figure 1 as well as Sections 2, 3 and 5. We provide high-quality panoramic 2D semantic annotations and 2D layout annotations and demonstrate its usability in the evaluation of three downstream supervised learning tasks: 3D semantic segmentation based on RGB-D images or point clouds and layout

[1] Ricoh Company, Ltd. Japan,
firstname.lastname@jp.ricoh.com
[2] Ricoh International B.V. - Niederlassung Deutschland, Germany,
firstname.lastname@ricoh-europe.com
[3] German Research Center for Artificial Intelligence, DFKI, Germany,
firstname.lastname@dfki.de
[4] RPTU Kaiserslautern, Germany
⋆ Denotes equal author contribution
This work was partially funded by the EU Horizon Europe Framework Program under GA 101058236 (HumanTech), and by the German Ministry for Economics and Climate Action (BMWK) under Grant 13IK010 (TWIN4TRUCKS).

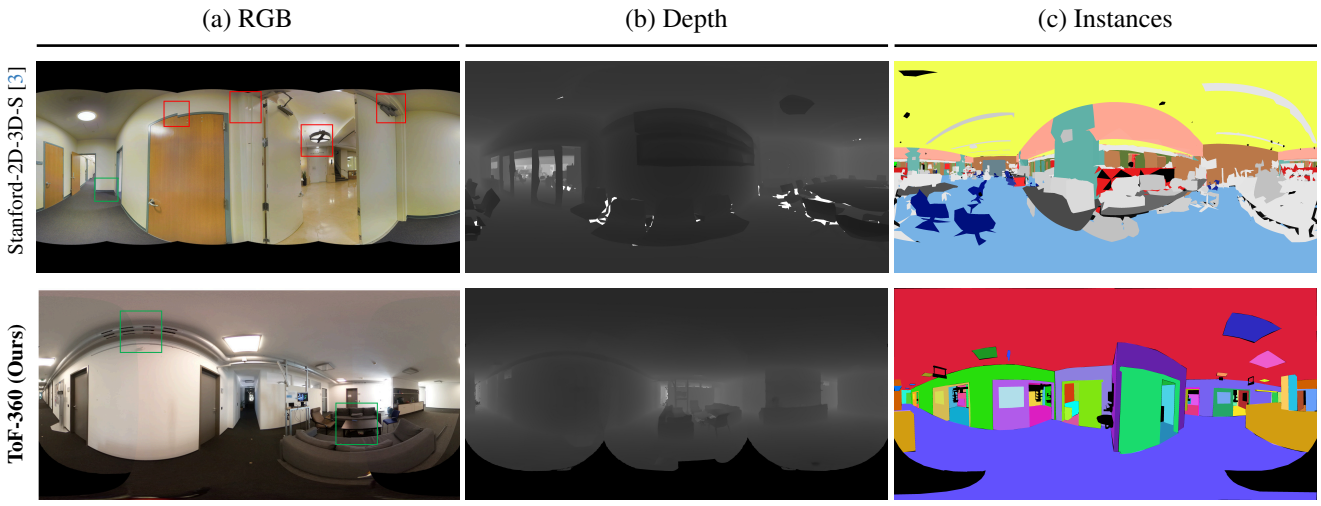

Figure 1. Comparison of multiple samples from Stanford-2D-3D-S [3] (**top**) and ToF-360 (**bottom**) showing superior quality stitching, depth, and instance labels in our ToF-360 dataset. (a) Comparison of image and stitching quality. Green boxes mark properly aligned stitching, while red boxes show misaligned stitches. (b) Qualitative depth comparison shows better depth edges and finer depth in our dataset. (c) Comparison of instance labels shows better label to object alignment.

estimation. For these tasks, ToF-360 provides the only RGB-D dataset labeled with building-defining object categories and image based layout boundaries (ceiling-wall, wall-floor) and its 3D structure, which are described in Section 4. Finally, we evaluate the performance of state-of-the-art methods when evaluated on ToF-360 in Section 6 and emphasize the challenges they face in generalization to domains unseen in training data. In summary, the contributions of this paper are as follows:

- We provide the first dataset created using an omnidirectional one-shot ToF device, the only scanner that can obtain omnidirectional distance information within a second.
- We provide high-quality hand-crafted segmentation and layout labels free of alignment and 3D-to-image label projection artifacts.
- We perform a comprehensive task evaluation in semantic segmentation using different modalities such as panoramic image based and point cloud based approaches.
- We introduce a benchmark for scene understanding tasks based on single-shot reconstruction without the need for global alignment and set a baseline using state-of-the-art methods.

2. Related Work

Machine learning algorithms for scene understanding are an active research area of great interest in computer vision, graphics, and robotics, and there is a growing demand for the collection of RGB-D images for training and evaluating such algorithms [17]. A comparison of ToF-360 to commonly used existing datasets is shown in Table 1. SUNCG [41] and Structured3D [54] provide large synthetic 3D datasets. While these can assume ideal conditions with a virtual renderer and can therefore construct diverse datasets with low noise, it is very difficult to approximate the quality of lighting and textures of the real world, which can lead to scene understanding conditions that deviate significantly from reality for some tasks. In recent years, several datasets have been published that capture real-world spaces with a general-purpose 3D scanner [6, 14]. Most of these datasets were captured by scanning with stationary terrestrial laser scanners (TLS) or handheld devices. Another ToF dataset, TIMo [38], combined infrared (IR) and depth for building monitoring and anomaly detection, while this dataset prioritizes privacy, it provides fewer annotations and modalities.

Our dataset, on the other hand, differs from the existing ones, providing 2D semantic annotations and room layouts in addition to 360° panoramic RGB images and panoramic depth images derived from ToF sensors. With this completely new capturing method and hardware and high data quality (see Figure 1), we contribute to the advancement of machine learning models for real-world tasks.

2.1. Panoramic Image Segmentation

Early methods for interpreting a picture holistically focused on using perspective image based models in conjunction with distorted-mitigated wide field of view (FoV) images. A distortion-mitigated locally-planar image grid tangent to a subdivided icosahedron is proposed by Eder et al. [16] for a tangent image spherical representation. Lee et al. [29], on the other hand, use a spherical polyhedron to symbolize comparable omni-directional perspectives. In contrast to that, recent studies [34] use distortion-aware modules in the network architecture to directly operate on equirectangular

representations. Sun et al. [42] suggest a discrete transformation for predicting dense features after an effective height compression module for latent feature representation. To improve the receptive field and learn the distortion distribution beforehand, Zheng et al. [55] combine the complementary horizontal and vertical representation in the same line of research. In an encoder-decoder framework, Shen et al. [39] introduced a brand-new panoramic transformer block to take the place of the convolutional block. Modern panoramic distortion-aware and deformable modules [13] have been added to the state-of-the-art UNet [37] and SegFormer [47] segmentation architectures to improve their performance in the spherical domain [19, 20, 34, 49, 50]. Making use of cross-modal interactions and panoramic perception abilities, SFSS-MMSI [20] jointly used the information from RGB-Depth-Normals modalities of equirectangular images and achieved state-of-the-art mIoU performance.

2.2. Point Cloud Semantic Segmentation

There are three main approaches for learning from 3D point clouds: projection-based, voxel-based, and point-based networks. **Projection-based** networks project point clouds onto regular grids and then process them with 2D convolutional neural networks (CNNs). This approach is intuitive but does not efficiently utilise the sparsity of point clouds and leads to loss in geometric information [7, 27]. **Voxel-based** networks convert point clouds into 3D voxels and then apply 3D convolutions. Those networks are computationally expensive and result in the loss of geometric detail due to quantisation [9, 18]. **Point-based** networks process point clouds directly as sets using permutation-invariant operators. They are more flexible and can better capture the geometric relationships between points. Some recent work has focused on using self-attention mechanisms in point-based networks, which has shown promise for large-scale 3D scene understanding [35, 36, 44]. Point Transformer by Zhao et al. [53] builds upon the foundations of point-based networks and self-attention mechanisms, utilising local self-attention [43], vector attention [52], and appropriate positional encoding. Point Transformer ushered the beginning of using transformers for semantic segmentation of point clouds as recent works have improved upon it to achieve state-of-the-art results [45, 46].

2.3. Room Layout Estimation

Room layout estimation is an important task in the process of 3D reconstruction and augmented reality (AR) applications aiming to estimate the boundaries of ceiling, floor, and walls [28]. As research interest in this task has grown, various datasets have emerged. Many existing public datasets (e.g. PanoContext [51] and LayoutNet [56]) assume a simple box layout for a single room. Matterport layout [57] extends to room layouts according to the Manhattan world assumption. Structured3D [54] provides more accurate room layouts based on a designed house model. Our dataset is based on images taken in the real world and annotated according to the Manhattan assumption, but unlike other datasets, it includes extensive public spaces such as offices and hospitals, which have a more complex structure than a typical room layout. This unique characteristic can be helpful in improving the robustness of layout models in real-world applications.

3. Omnidirectional ToF RGB-D Device

3.1. Hardware Configuration

The used 3D spatial sensing device is depicted in Figure 2. The device's upper portion includes two built-in fisheye RICOH THETA [30] cameras with more than $180°$ FoV each for capturing omnidirectional RGB images. Furthermore, ToF LiDAR emitters and detectors are installed for the collection of $360°$ depth information. In more detail, two fisheye lenses that provide RGB information and four fisheye lenses that gather ToF depth information are used to create the omnidirectional image. The circuit board for processing the acquired data, the battery for the processing system, and the ToF laser emitter are all located in the lower portion of the device. The depth information is aligned with the RGB images using calibration parameters provided from the device assembly.

3.2. Device Specifications

In Table 1 the device specifications are displayed and contrasted with those of currently available, widely-used 3D capturing devices. In contrast to these conventional scanners, we are able to capture the entire space with a brief light exposure by using ToF and uniform illumination. The acquisition speed is nearly $12\times$ faster than the Matterport Pro2 [23] and the depth resolution is larger than any other scanners in the table. High frame rates can be reached with the portable scanners, however, the point measurement rate is constrained by the FoV.

4. Dataset Details

In this section, we describe the steps followed to acquire our ToF-360 dataset from raw data collection in real-world buildings as well as the manual annotation of semantic labels, and room layout annotation process.

4.1. RGB-D Panoramas

ToF-360 contains 5792×2896 (width$\times$height) color, depth and XYZ (coordinate) equirectangular images covering approximately 360×300 degrees (horizontal$\times$vertical, the entire sphere except the bottom area). 207 total panoramas were collected from real buildings that contain 4 scenes from an office building, car parking area, and an empty hospital. The scenes are broken down as: 40 and 44 panoramas from

Table 1. Comparison of 3D scene datasets. ToF-360 provides the widest field of view, the highest density 360° RGB-D images, and has the highest scanning speed of any panoramic depth sensor. Depth resolution and field-of-view in this table are indicated as width (horizontal) × height (vertical). The first section presents captured dataset while the second section presents synthetic datasets.

Datasets	Size	Classes	Sensing type	Depth resolution	Field-of-View	Ideal Range	FPS
SUN RGB-D [40]	10,335 images 47 scenes	800	Sequential	628 × 468 320 × 240 512 × 424 640 × 480	87° × 58° 57° × 43° 70° × 60° 58° × 45°	0.6-3.5 m 0.8-4 m 0.5-4.5 m 0.8-3.5 m	90 30 30 30
Stanford-2D-3D-S [3]	1413 images 11 scenes	13	Panorama	4096 × 2048	360° × 300°	0.5-5 m	0.04
ScanNet [12]	1,513 scans 707 venues	40	Sequential	640 × 480	58° × 45°	0.4-3.5 m	60
Matterport3D [6]	2,056 scans 2,056 rooms	40	Panorama	2048 × 1024	360° × 300°	0.5-5 m	0.04
ARKitScenes [14]	5048 scans 1661 scenes	17	Sequential Panorama	256 × 192 1920 × 1440	122° × 122° 360° × 300°	0.5-5 m 0.6-70 m	60 0.007
SUNCG [41]	404,058 rooms 45,622 scenes	84	Sequential	640 × 480	N/A	N/A	N/A
Structured3D [54]	196,515 frames 3,500 scenes	40	Panorama	512 × 1024 720 × 1280	N/A	N/A	N/A
ToF-360 (Ours)	207 images 4 scenes	39	Panorama	5792 × 2896	360° × 300°	0.5-5 m	0.5

two office floors, 43 panoramas from the parking lot, and 52 panoramas from the empty hospital floor. Various examples of the different scenes are presented in the supplementary material.

4.2. Raw Data Acquisition Process

We used our device described in Section 3. The data acquisition process uses a tripod-mounted device in a fixed orientation relative to the scene at approximately the height of a human observer. All the personally identifiable information such as the nameplate in the office area and number plates in the parking area was blurred manually by the annotators after data recordings. Figure 2 depicts the workflow for creating panoramic images from the RGB-D photos that the device has collected. Four fisheye depth images are created from the ToF raw data from the LiDAR component. Since the RGB spatial resolution is higher than that of the LiDAR, upsampling processing based on nearest neighbor search [4] and bilinear interpolation is used to fill in the missing depth region. The RGB data and the collection of distance measurements are aligned by the intrinsic and extrinsic parameters that were calculated by a checkerboard calibration using OpenCV [24]. The conversion of fisheye images to equivalent ERP representation is inspired by [15]. Since the device's baseline is around 6 cm between the RGB lenses, naively stitching the border between each lens results in occlusion along the device's center line. To avoid this, each RGB value captured by the first lens was converted into 3D coordinates with a corresponding depth value and projected towards the second lens' focal, and vice versa. As a result, our method creates a 360° panoramic RGB–Depth image of the scene with minimal artifacts.

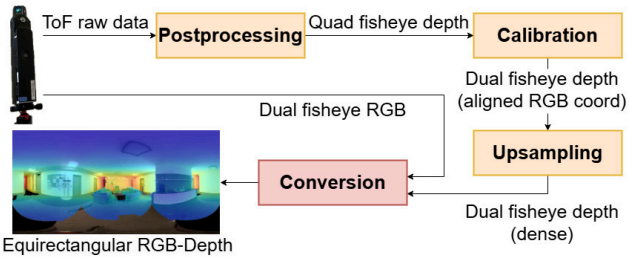

Figure 2. Our data acquisition pipeline. The quad-fisheye depth images obtained from the device are converted to a dual-fisheye depth image aligned with the dual-fisheye RGB image by calibration. RGB-D dual-fisheye images are converted to an equirectangular image.

4.3. Semantic Annotation

For the semantic annotation of the data, (see Section 4.1), we used the COCO Annotator [5] for labelling the RGB data. We follow ontology-based annotation guidelines developed for both RGB-D and point cloud data [26]. These guidelines take into consideration the differences between both image and point cloud modalities. Due to the unpredictable ways a depth sensor can interact with glass, both the glass surfaces (e.g. windows, doors) and the objects behind them are annotated.

4.4. Layout Annotation

We used PanoAnnotator [48] as an annotation tool for the room layout. All inputs are preprocessed by function sets in PanoAnnotator to generate Manhattan-aligned panoramas. This deformation is based on a line-detection algorithm and panorama rotation following [51] and its rotation matrix to ensure compatibility with the original image. Each layout element (ceiling, wall, floor, openings) is manually annotated and stored in json file format which contains the position of layout corner points, and the plane equation of each layout element in the 3D world. Elements occluded from the original acquisition point have also been annotated by following the actual building geometry as far as possible.

5. Data Quality

In this section, we compare the quality among our ToF-360 and other datasets for the four modalities described in Section 4 - RGB, depth, instances, and room layout.

5.1. RGB images

The main datasets providing panoramic images use the Matterport device [3, 6]. It uses three cameras rotated in six directions to obtain a 360° panorama, with stitching between images occurring in six horizontal and three vertical locations. In contrast, our device employs two hemispheric cameras, so stitching between images occurs in only two places in the horizontal direction. The quality of the stitching lines is influenced by the calibration accuracy between the camera lenses and the interpolation algorithm. We use depth information for the 3D projection of the RGB of each lens and then convert it to a 2D representation by binocular integration, which results in less distortion in the stitching lines and data with good pixel correspondence between RGB and depth. For qualitative differences, see Figure 1 and the supplementary material.

5.2. Depth images

A 360° RGB-D dataset similar to our conditions is Stanford-2D-3D-S [3]. Their depth images are generated by rendering the reconstructed 3D meshes from the camera viewpoints. In contrast, our ToF-360 provides depth images without any back projection from reconstructed 3D meshes. More specifically, instead of constructing the panoramic depth from multi-viewpoint measurements, a depth image is obtained independently for each recording point. This is advantageous for our dataset in terms of recording time, expertise required by the person recording, and time and specialized software needed for post-processing the acquired data. Additionally, this can make the removal of dynamic objects in post-processing much simpler and reduce the loss of information.

5.3. Instances

Our instance annotation is done manually by annotators directly on the images. The instances we provide are annotated on every pixel. In contrast, the masks in the Stanford-2D-3D-S [3] dataset are generated by projecting the labels from the 3D meshes onto the 2D images which leads to visible artifacts as seen in Figure 1.

5.4. Room Layout

Our ToF-360 provides a more complex room geometry rather than the simple cuboid room layout provided by Stanford-2D-3D-S [3] and PanoContext [51]. MatterportLayout [57] assumes the Manhattan hypothesis, which is very similar to our setup but sometimes contain annotations that do not reflect the actual room layout. Figure 3 shows a sample of qualitative results: the MatterportLayout [57] example has one less annotated plane, which generates unnatural layout boundaries. We paid close attention to annotations where the layout boundaries match the real building structure.

6. Evaluation

Our evaluation is primarily meant to demonstrate the challenges of cross-dataset adaptation of existing scene understanding models. Therefore, we evaluate the generalization capabilities of state-of-the-art models trained on public datasets [2, 3, 6, 12, 51, 54] and then tested on our ToF-360 dataset. The evaluation is done on the semantic segmentation task for both image based, and point cloud based semantic segmentation as well as for the layout estimation task.

6.1. Task: Semantic Segmentation

The semantic segmentation evaluation of image and point cloud based methods is presented in Table 2. For the image based semantic segmentation, we resize the input image to 512×1024. We compute evaluation metrics, such as Mean Region Intersection Over Union ($mIoU$), Pixel Accuracy ($aAcc$), and Mean Accuracy ($mAcc$), using the *MMSegmentation* scripts[1]. The current state-of-the-art approaches: HoHoNet [42], PanoFormer [39], and SFSS-MMSI [20] are used for image based RGB+(D) segmentation experiments. For a detailed description of their implementation details, please refer to the corresponding works. The models are trained on the **Stanford-2D-3D-S** [3] and **Structured3D** [54] datasets and are then evaluated on our ToF-360 dataset.

The **Stanford-2D-3D-S** [3] dataset consists of multi-modal equirectangular images with 13 object categories and divided into 6 Areas. We use the `fold_1` split for training and validation as suggested by Armeni et al. [3]. The **Structured3D** [54] dataset is a synthetic dataset that offers

[1] https://mmsegmentation.readthedocs.io/en/0.x/

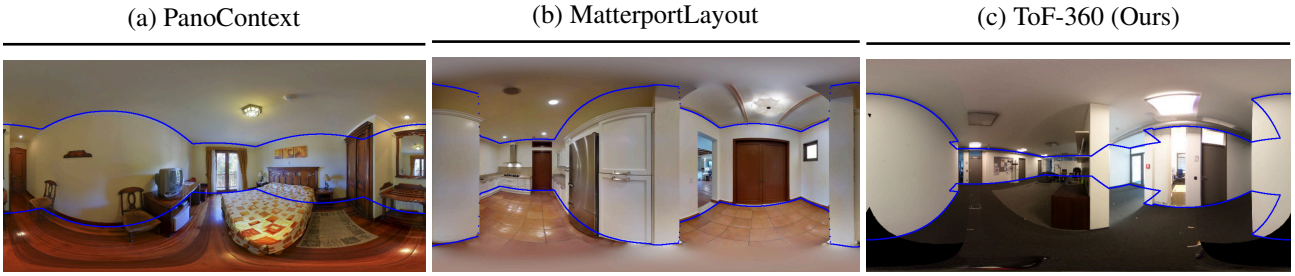

Figure 3. Examples of annotation failure on (a) PanoContext [51] and (b) MatterportLayout [57]. In contrast, (c) ToF-360 provides correct room layout annotations. Ground truth boundary is shown as blue line.

40 NYU-Depth-v2 [21] object categories and multi-modal, equirectangular images with a variety of lighting setups. We use the train, validation, and test splits as described by Zheng et al. [54]. The best validation performance checkpoints of respective models are reported.

As a pre-processing step, the object semantics from our proposed panoramic dataset (see Section 4.1 and Section 4.3) are respectively remapped to 13 object categories for Stanford-2D-3D-S [3] and 40 NYU-Depth-v2 [21] object categories for Structured3D [54] dataset experiments. The mapping is provided with the dataset for the reproducibility of the results.

In addition to the image based semantic segmentation, we evaluated on our dataset a state-of-the-art point cloud segmentation model [46] trained on existing public datasets. The single-scan inputs are first voxel downsampled to 1 $point/cm$ then evaluated using the Point Transformer V3 (PTv3) model [46]. The model was trained on a joint dataset comprising of **ScanNet** [12], **Structured3D** [54], and **S3DIS** [2] datasets and validated on the **S3DIS** validation set. The training setup is described in detail by Wu et al. [46] and replicated for this evaluation. The model achieving the best validation result was chosen for the evaluation of the ToF-360 point clouds. The **S3DIS** [2] dataset is the point cloud version of **Stanford-2D-3D-S** [3] and follows the same structure and number of classes.

Similar to the image based semantic segmentation, we also compute the evaluation using the $mIoU$, $mAcc$ and $aAcc$ metrics however over the points instead of pixels. We used the coordinates, color, and normals as input modalities and used the same category mapping from our proposed dataset categories to respective 13 object categories as done in the image based evaluation.

We carry out comprehensive tests on the proposed RGB-Depth-Normals panoramic ToF dataset from real-world setting. Figure 4 and Figure 5 present the qualitative results of the evaluation of the image based and point cloud based segmentation evaluations on the Stanford-2D-3D-S and S3DIS datasets respectively, and Table 2 presents the quantitative results for both image based and point cloud based segmentation approaches.

Table 2. Evaluation of semantic segmentation performance for the proposed ToF-360 dataset trained on Stanford-2D-3D-S [3] for image based approaches and S3DIS [2] for the point cloud based approach.

Method	**Modalities**	**Results**		
		$mIoU(\%)$	$mAcc(\%)$	$aAcc(\%)$
HoHoNet [42]	RGB	20.76	41.65	66.90
PanoFormer [39]		28.07	51.44	**77.75**
SFSS-MMSI [20]		**29.56**	**51.53**	76.65
HoHoNet [42]	RGB-D	**27.46**	**48.66**	**76.11**
PanoFormer [39]		21.52	39.90	65.30
SFSS-MMSI [20]		24.92	45.88	73.11
SFSS-MMSI [20]	RGB-D-N	**23.17**	**46.26**	**70.39**
PTv3 [46]	RGB-XYZ-N	18.57	25.08	67.89

The results in Table 2 are better for the image based approaches compared to the point based approach. This can be attributed to the higher similarity in the data representation in the RGB domain (larger domain gap for point clouds). While both the train and test data are equirectangular RGB-D images for the image-based approaches, the point cloud approach was trained on more complete point clouds unlike the single shot point clouds generated by our sensor due to single view occlusions. This means that the image based methods are more capable of generalizing when applied to ToF-360.

Another challenge to the generalization of the models is caused by the differences in recorded areas as well as labeled classes. Unlike the Stanford-2D-3D-S and S3DIS datasets which are predominantly recorded in office areas, our dataset includes scenes from new and challenging settings (parking lot and hospital).

Figure 5 qualitatively demonstrates that good results are obtained on structural objects such as walls, floor, and ceiling while the other classes are mostly detected as clutter.

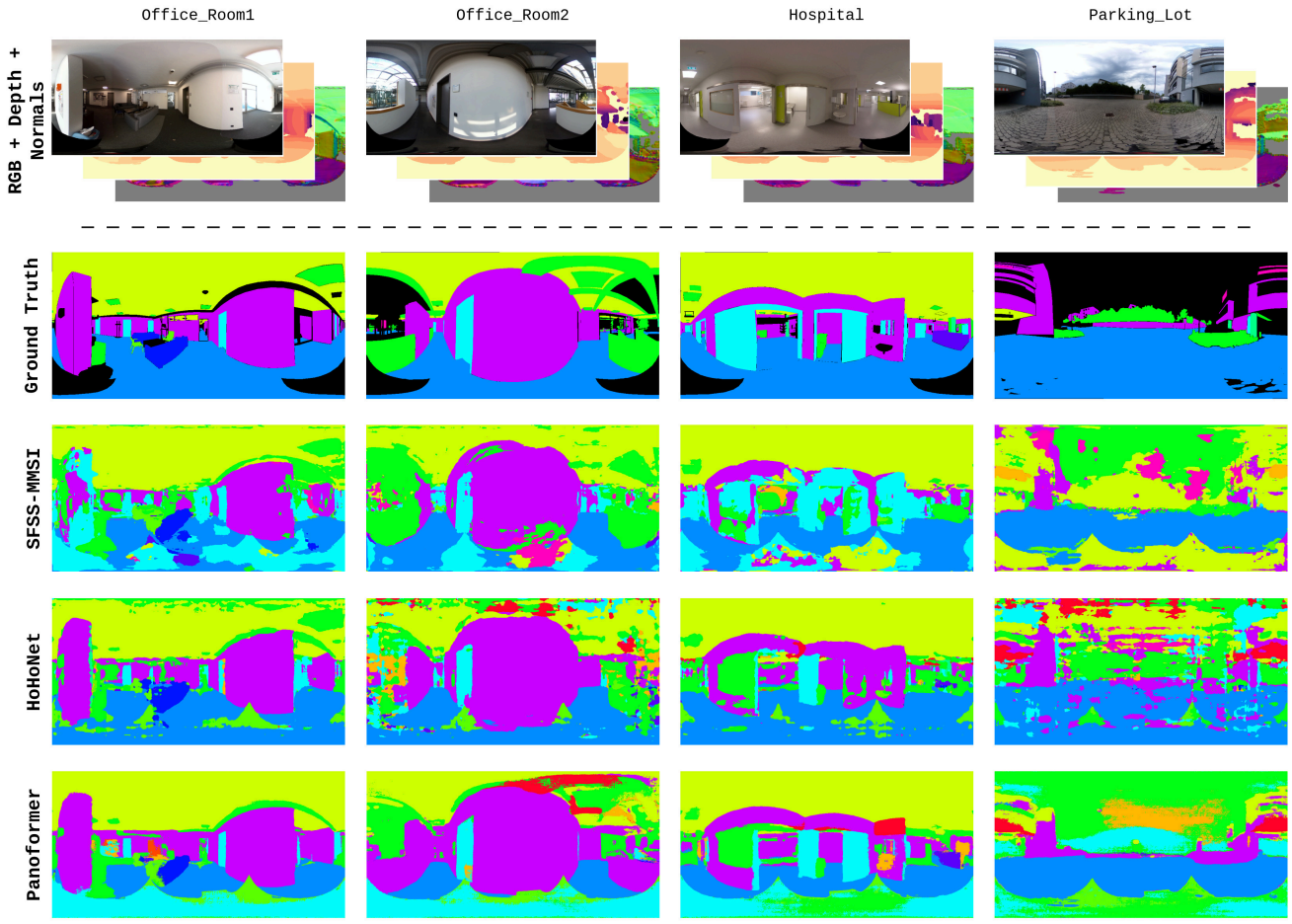

Figure 4. Visualization of RGB-Depth-Normals semantic segmentation results for the proposed ToF-360 dataset. In the above visualization, SFSS-MMSI [20] is trained with RGB-Depth + Normals while HoHoNet [42] and PanoFormer [39] with RGB-Depth panoramic equirectangular images from Stanford-2D-3D-S [3].

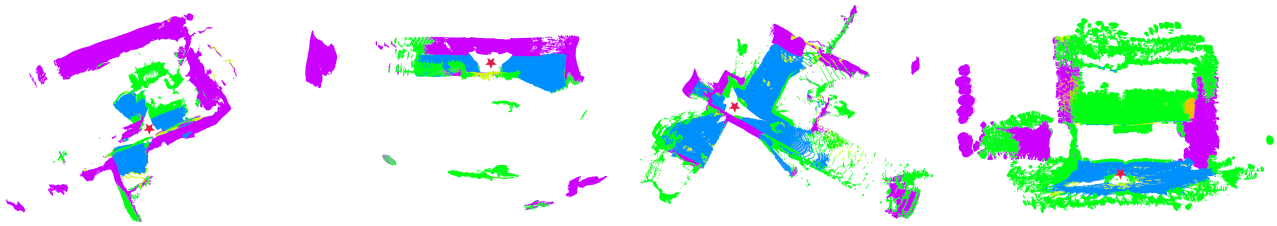

Figure 5. Visualization of the results of the point cloud based semantic segmentation using Point Transformer V3 by Wu et al. [46], the colors correspond to the same classes as in Figure 4. The ceiling has been removed for the indoor scenes (`Office_Room1`, `Office_Room2`, and `Hospital`) due to visualization limitations. The outdoor scene is showing the side of the building (cropped in the center in RGB) for easier understanding. The location of the sensor during recording is marked with a red star ⋆.

This further supports the argument that incomplete scans (single-shot) lead to lower detection accuracy on some objects such as the furniture and doors. When comparing the results from Figure 4 and Figure 5 we can see that the image based approaches generalized better on the furniture classes and both approaches performed similarly for the structural elements.

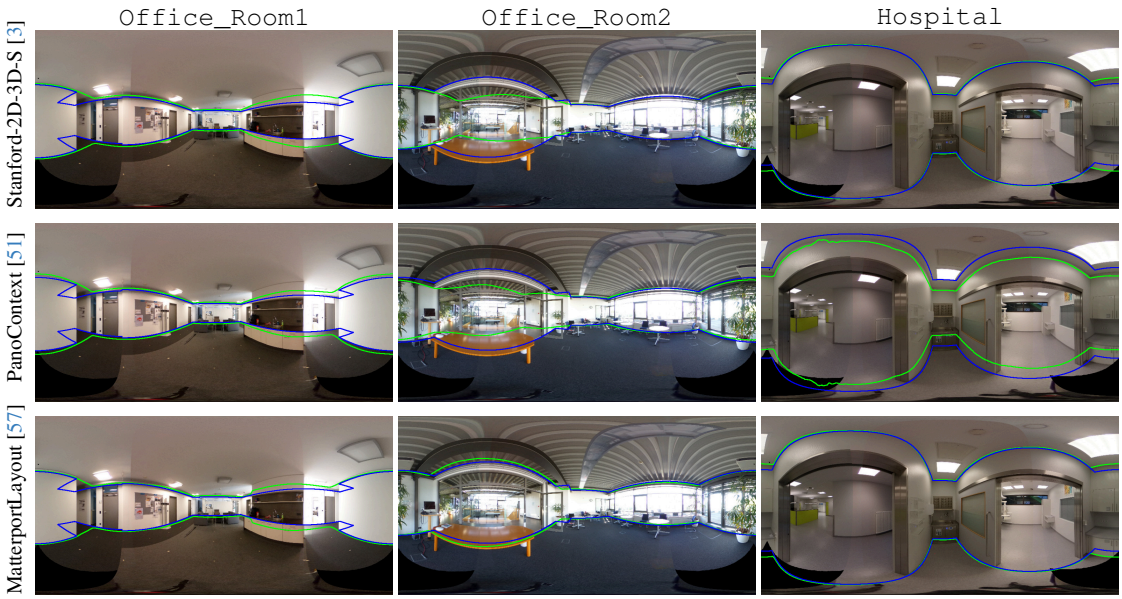

Figure 6. Qualitative comparison of layout estimation methods on ToF-360 produced using LGT-Net[25]. We show the boundaries of room layout on panorama. The blue lines are ground truth, and the green lines are prediction.

6.2. Task: Layout Estimation

We present the evaluation of our dataset ToF-360 for the room layout estimation task in Table 3. The input images are resized to 512×1024, and standard evaluation metrics including intersection over union of floor shapes ($2DIoU$) and 3D room layouts ($3DIoU$), root mean squared error ($RMSE$) of estimated depth, and the ratio between prediction depth and ground truth depth within threshold of 1.25 (δ_1) are calculated following [56]. We used the layout estimation models provided by LGTNet [25]. These models are pre-trained by the authors with public datasets consisting of Stanford-2D-3D-S [3], PanoContext [51], and MatterportLayout [6]. Two images from the hospital scene and all of the parking lot scene were removed for layout estimation since they do not adhere to the Manhattan assumption.

We perform tests on our proposed panoramic image dataset recorded in the real world. We show the quantitative evaluation in Table 3 and qualitative comparisons in Figure 6. The results on MatterportLayout are better than others. The scenes provided by ToF-360 sometimes have large openings in the walls, such as windows and doors, as shown in the `Office_Room2` and `Hospital` results. The quantitative results of our proposed ToF-360 are supported by the fact that the Stanford-2D-3D-S and PanoContext only offer cuboid room layouts, whereas the MatterportLayout provides a Manhattan-aligned structure.

Table 3. Quantitative results of layout estimation methods on ToF-360 produced using LGT-Net [25]. IoU values are in %, for $RMSE$ lower values are better.

Trained dataset	$2DIoU^{\uparrow}$	$3DIoU^{\uparrow}$	$RMSE^{\downarrow}$	$\delta_1^{\uparrow}$
Stanford-2D-3D-S [3]	59.66	57.33	0.742	0.831
PanoContext [51]	60.20	57.80	0.770	0.849
MatterportLayout [57]	**62.71**	**62.88**	**0.730**	**0.900**

7. Conclusion

We introduced ToF-360, a unique RGB-D dataset created using an omnidirectional one-shot ToF device, the only scanner that can obtain 360-degree distance information in one second. We provide instance-level semantic annotations labeled with building-defining object categories and image based layout boundaries. We proposed a comprehensive evaluation for semantic segmentation using different modalities such as panoramic image based and point cloud based approaches. This defines a new benchmark for single-shot reconstruction without the need for global alignment. As our dataset is confined to a limited number of real-world scenes, it serves as a challenge to existing works and shows the difficulties in generalization of models training on existing datasets especially with regards to non-uniformity of annotation and scene bias. Future work is planned to extend and generalize the dataset through continuous data acquisition and additional annotations as well as investigate methods to reduce the domain gap between 3D semantic segmentation datasets.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv e-prints*, abs/1812.11941:arXiv:1812.11941, 2018. 1

[2] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, 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. 5, 6

[3] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2, 4, 5, 6, 7, 8, 1

[4] Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *JACM*, 1998. 4

[5] Justin Brooks. COCO Annotator. https://github.com/jsbroks/coco-annotator/, 2019. 4

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2, 4, 5, 8

[7] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 3

[8] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Understanding indoor scenes using 3d geometric phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 33–40, 2013. 1

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 3

[10] Microsoft Cooporation. Microsoft kinect v2. https://learn.microsoft.com/en-us/azure/kinect-dk/windows-comparison#hardware, 2024. 1

[11] Intel Corporation. Intel® realsense™ lidar camera l515. https://ark.intel.com/content/www/us/en/ark/products/201775/intel-realsense-lidar-camera-l515.html, 2023. 1

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 4, 5, 6

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773. IEEE Computer Society, 2017. 3

[14] Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *NeurIPS Datasets and Benchmarks*, 2(6):16, 2021. 2, 4

[15] Xiaoming Deng, Fuchao Wu, Yihong Wu, and Chongwei Wan. Automatic spherical panorama generation with two fisheye images. In *2008 7th World Congress on Intelligent Control and Automation*, pages 5955–5959. IEEE, 2008. 4

[16] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *CVPR*, pages 12423–12431. Computer Vision Foundation / IEEE, 2020. 2

[17] Michael Firman. Rgbd datasets: Past, present and future. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 19–31, 2016. 2

[18] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 3

[19] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and Jose J Guerrero. What's in my room? object recognition on indoor panoramic images. In *IEEE ICRA*, 2020. 3

[20] Suresh Guttikonda and Jason R. Rambach. Single frame semantic segmentation using multi-modal spherical images. In *WACV*, pages 3210–3219. IEEE, 2024. 3, 5, 6, 7

[21] Dmitry Ignatov, Andrey Ignatov, and Radu Timofte. Virtually enriched nyu depth v2 dataset for monocular depth estimation: Do we need artificial augmentation? In *CVPR*, 2024. 6

[22] Apple Inc. iphone 13 pro. https://support.apple.com/kb/SP852?locale=en_US, 2023. 1

[23] Matterport Inc. Matterport pro2 3d camera. https://matterport.com/pro2, 2023. 3

[24] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015. 4

[25] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, 2022. 8

[26] Fabian Kaufmann, Mahdi Chamseddine, Suresh Guttikonda, Christian Glock, Didier Stricker, and Jason Rambach. Ontology-based semantic labeling for rgb-d and point cloud datasets. In *EC3 Conference 2023*, pages 0–0. European Council on Computing in Construction, 2023. 4

[27] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 3

[28] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 4865–4874, 2017. 3

[29] Yeon Kun Lee, Jaeseok Jeong, Jong Seob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *CVPR*, pages 9181–9189. Computer Vision Foundation / IEEE, 2019. 2

[30] Ricoh Company Ltd. Ricoh theta v camera. https://theta360.com/en/about/theta/v.html, 2023. 3

[31] Ricoh Company Ltd. Ricoh 3d reconstruction device. https://www.ricoh.com/technology/tech/126_building_digital_twin, 2024. 1

[32] Muzammal Naseer, Salman Khan, and Fatih Porikli. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access*, 7:1859–1887, 2018. 1

[33] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[34] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. *Signal Image Video Process.*, 16 (3):643–650, 2022. 2, 3

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, pages 234–241. Springer, 2015. 3

[38] Pascal Schneider, Yuriy Anisimov, Raisul Islam, Bruno Mirbach, Jason Rambach, Didier Stricker, and Frédéric Grandidier. Timo—a dataset for indoor building monitoring with a time-of-flight camera. *Sensors*, 22(11):3992, 2022. 2

[39] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360$ˆ{\circ }$ depth estimation. In *ECCV (1)*, pages 195–211. Springer, 2022. 3, 5, 6, 7

[40] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 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. 1, 4

[41] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2, 4

[42] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *IEEE/CVF CVPR*, 2021. 3, 5, 6, 7

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 3

[45] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 3

[46] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024. 3, 6, 7

[47] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, pages 12077–12090, 2021. 3

[48] Shang-Ta Yang, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. Panoannotator: a semi-automatic tool for indoor panorama layout annotation. In *SIGGRAPH ASIA Posters*, pages 34:1–34:2. ACM, 2018. 5

[49] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *CVPR*, pages 16896–16906. IEEE, 2022. 3

[50] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *CoRR*, abs/2207.11860, 2022. 3

[51] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *ECCV (6)*, pages 668–686. Springer, 2014. 3, 5, 6, 8

[52] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10076–10085, 2020. 3

[53] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 1, 3

[54] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 2, 3, 4, 5, 6

[55] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. Complementary bi-directional feature compression for indoor 360° semantic segmentation with self-distillation. In *WACV*, pages 4490–4499. IEEE, 2023. 3

[56] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *IEEE/CVF CVPR*, 2018. 3, 8

[57] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360° image: A comparative study of state-of-the-art methods. *arXiv preprint arXiv:1910.04099*, 2019. 3, 5, 6, 8