ToF-360：ワンショット撮影による屋内シーンの360°RGB-D セマンティック3D復元用データセット

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

金山　英誠[*]　マフディ　シャムスッディーン[**]　スレシュ　グッティコンダ[**]
Hideaki KANAYAMA　Mahdi CHAMSEDDINE　Suresh GUTTIKONDA

奥村　聡[*]　横田　聡一郎[*]　ディディエ　ストリッカー[**]　ジェイソン　ランバック[**]
So OKUMURA　Soichiro YOKOTA　Didier STRICKER　Jason RAMBACH

要　旨

3Dシーン理解は，自動化分野における重要な研究課題である．既存のRGB-Dデータセットはシーン全体の再構築に焦点を当てているが，取得には視野角の限られたスキャナによる複数回のスキャンが必要で，時間的負担が大きい．また，スティッチングによるアーティファクトや低品質なアノテーションも課題である．本研究では，360度全方位RGB-Dスキャンを数秒で実現する独自のTime-of-Flightセンサーを用いた「ToF-360」データセットを提案する．魚眼形式およびERP画像に加え，高精度な2Dセマンティクスと室内レイアウトのアノテーションを提供し，2D/3Dセグメンテーションおよびレイアウト推定のベンチマークを導入する．ToF-360は，実環境の表現力向上と既存手法の性能改善に貢献する．データセットは以下より公開されている．https://doi.org/10.57967/hf/5074

ABSTRACT

3D scene understanding is a critical research topic in automation for various automation areas. Existing RGB-D datasets focus on full-scene reconstruction but require time-consuming multi-pass scanning with limited field-of-view devices, often resulting in stitching artifacts and low-quality annotations. We present ToF-360, the first RGB-D dataset captured using a unique Time-of-Flight sensor enabling 360° omnidirectional scanning within seconds. The dataset includes fisheye and ERP images, along with manually annotated pixel-level 2D semantics and room layouts. Benchmarks are provided for 2D semantic segmentation, 3D semantic segmentation, and layout estimation. ToF-360 enhances real-world representation and improves the performance of state-of-the-art methods. The dataset is publicly available at https://doi.org/10.57967/hf/5074

*　技術統括部　先端技術研究所
Advanced Technology Research Institute, Technology Management Division

**　ドイツ人工知能研究所 (DFKI)
German Research Center for Artificial Intelligence (DFKI)

本稿を引用する際は，以下の学術会議にて発表された元の論文を参照すること．
For citation purposes, please refer to the original publication presented at the following conference:
Kanayama, H., Chamseddine, M., Guttikonda, S., Okumura, S., Yokota, S., Stricker, D., & Rambach, J. (2025). ToF-360: A Panoramic Time-of-Flight RGB-D Dataset for Single Capture Indoor Semantic 3D Reconstruction. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 4442–4451).

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[1,2]. Many recent machine learning techniques including depth estimation, 3D reconstruction, and semantic segmentation[3-5] 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 a dataset such as Matterport3D[6] exhibits image alignment artifacts or low-quality segmentation mask labels directly projected from 3D (see Fig. 1).

Even though mobile scanning LiDAR devices like Microsoft Kinect[7], iPhone LiDAR[8], and the RealSense LiDAR camera[9] facilitate data collection, continuous scanning is required because of the restriction on the angle of the view angle to <120°, which extends the acquisition time. These limitations while recording data can be a barrier to adoption to practical applications.

Ricoh released a novel Time-of-Flight (ToF) device capable of capturing complete colored 3D point clouds omnidirectionally from a single camera shot[10]. The device employs an indirect ToF method using amplitude-modulated infrared light, providing a horizontal field of view of 360° and a vertical field of view of 150°, with approximately 30° missing on the bottom side.

It solves the 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. In methods that integrate measurements from multiple viewpoints to generate per-view depth images, it is a prerequisite that the resulting 3D mesh achieves sufficient geometric accuracy. In contrast, single-shot measurements do not require multi-view integration, which avoids error propagation caused by inaccurate camera pose estimation across multiple views.

Our dataset, ToF-360, consists of 179 spherical RGB-D images taken in four unique environments. We emphasize the precise annotation and superior data quality of our dataset, compared to other datasets and 3D scanners in Fig. 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 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.

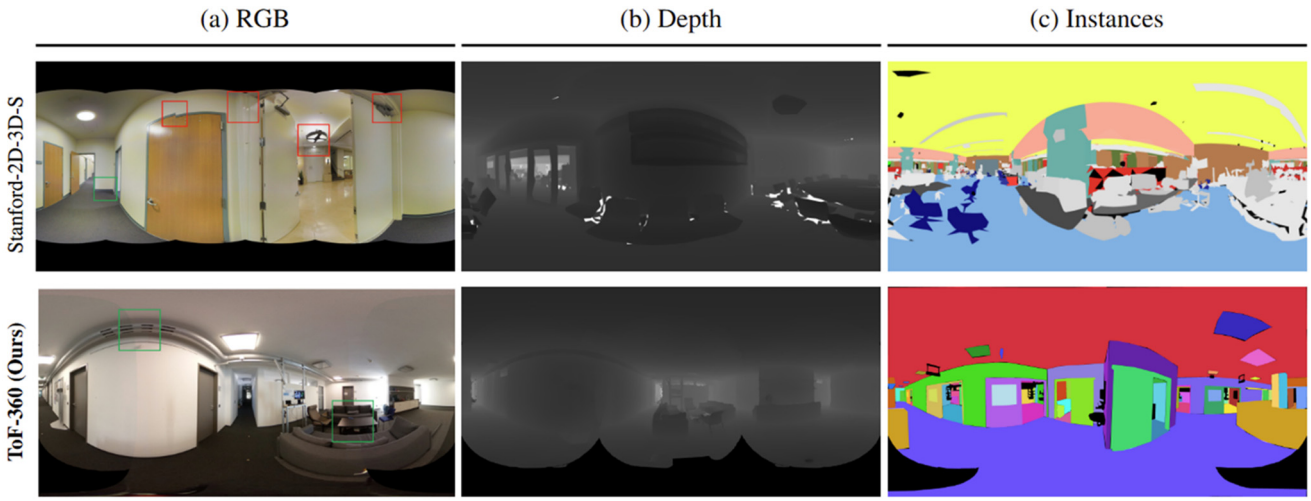

Fig. 1 Comparison of multiple samples from Stanford-2D-3D-S (top) and ToF-360 (bottom) showing superior quality stitching, depth, and instance labels in our ToF-360 dataset. (a) Comparison of image and stitching quality. Green boxes mark properly aligned stitching, while red boxes show misaligned stitches. (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.

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[11]. A comparison of ToF-360 to commonly used existing datasets is shown in Table 1. SUNCG[12] and Structured3D[13] 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[15], 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 Fig. 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 distortion-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.[17], on the other hand, use a spherical polyhedron to symbolize comparable omnidirectional perspectives. In contrast to that, recent studies[18] use distortion-aware modules in the network architecture to directly operate on equirectangular

representations. Sun et al.[19] 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.[20] combine the complementary horizontal and vertical representation in the same line of research. In an encoder-decoder framework, Shen et al.[21] introduced a brand-new panoramic transformer block to take the place of the convolutional block. Modern panoramic distortion-aware and deformable modules[22] have been added to the state-of-the-art UNet[23] and SegFormer[24] segmentation architectures to improve their performance in the spherical domain[18,25-28]. Making use of cross-modal interactions and panoramic perception abilities, SFSS-MMSI[26] 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 utilize the sparsity of point clouds and leads to loss in geometric information[29,30]. 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[31,32]. 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[33-35]. Point Transformer by Zhao et al.[5] builds upon the foundations of point-based networks and self-attention mechanisms, utilizing local self-attention[36], vector attention[37], 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[38,39].

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[40]. As research interest in this task has grown, various datasets have emerged. Many existing public datasets (e.g., PanoContext[41] and LayoutNet[42]) assume a simple box layout for a single room. Matterport layout[43] extends to room layouts according to the Manhattan world assumption. Structured3D[13] 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 Fig. 2. The device's upper portion includes two built-in fisheye RICOH THETA[44] 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.

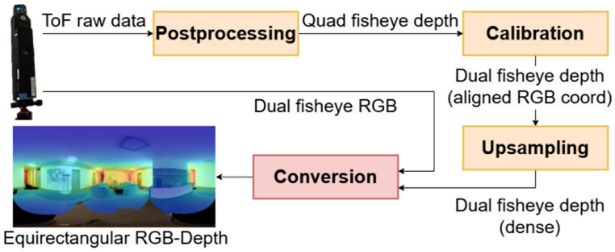

Fig. 2 Our data acquisition pipeline. The quad-fisheye depth images obtained from the device are converted to a dual-fisheye depth image aligned with the dual-fisheye RGB image by calibration. RGB-D dual-fisheye images are converted to an equirectangular image.

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 depth resolution is larger than any other scanners in the table. Reducing the resolution gap between RGB and depth modalities facilitates accurate pixel-level correspondence, enabling more precise integration of color and geometric information for tasks such as segmentation. High frame rates can be reached with the portable scanners, however, the point measurement rate is constrained by the FoV.

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[45]	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[52]	1413 images 11 scenes	13	Panorama	4096 × 2048	360° × 300°	0.5-5 m	0.04
ScanNet[53]	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[12]	404,058 rooms 45,622 scenes	84	Sequential	640 × 480	N/A	N/A	N/A
Structured3D[13]	196,515 frames 3,500 scenes	40	Panorama	512 × 1024 720 × 1280	N/A	N/A	N/A
ToF-360 (Ours)	179 images 4 scenes	39	Panorama	5792 × 2896	360° × 300°	0.5-5 m	0.5

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 × height) color, depth, and XYZ (coordinate) equirectangular images covering approximately 360 × 300 degrees (horizontal × vertical, the entire sphere except the bottom area). A total of 179 panoramas were collected from real buildings, including an office building, a car parking area, and an empty hospital. Unlike existing datasets that mainly cover simple residential spaces, these environments were chosen to provide structural and material diversity for both layout estimation and semantic segmentation. Offices feature long corridors, partition walls, and transparent glass walls; parking areas provide large open spaces with pillars; hospitals include complex room connections. This variety helps evaluate the robustness of models trained on existing datasets in real-world scenarios. The scenes are broken down as: 40 and 44 panoramas from two office floors, 43 panoramas from the parking lot, and 52 panoramas from the empty hospital floor.

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 personally identifiable information such as nameplates in the office area and number plates in the parking area was blurred manually by the annotators after data recordings. Fig. 2 depicts the workflow for creating panoramic images from the RGB-D photos collected by the device. 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[46] and bilinear interpolation is used to fill in the missing depth regions. Additionally, intensity edges are extracted from RGB images using OpenCV's Structured Edge Detection (SED)[47], and depth discontinuities are detected by evaluating local depth variation along these edges. If the depth difference within the neighborhood exceeds 100 mm, the corresponding edge region is excluded from depth completion to avoid propagating large depth errors. The RGB data and the collection of distance measurements are aligned by the intrinsic and extrinsic parameters calculated by a checkerboard calibration using OpenCV[47]. The conversion of fisheye images to equivalent ERP representation is inspired by[48]. Because the two fisheye lenses have a baseline of approximately 6 cm, directly stitching their borders introduces occlusions along the center line. To address this, one fisheye image is kept as-is for half of the equirectangular panorama, while the other image is first back-projected into 3D using its associated depth values. These 3D points are then reprojected onto a virtual fisheye camera with an idealized zero-baseline configuration, ensuring that overlapping regions are geometrically consistent and minimizing parallax artifacts before stitching. In the overlapping regions of both images, the pixel with the higher RGB intensity is selected, and the corresponding depth value is assigned using the same rule.

4-3 Semantic Annotation

For the semantic annotation of the data (see Section 4-1), we used the COCO Annotator[49] for labeling the RGB data. We follow ontology-based annotation guidelines developed for both RGB-D and point cloud data[50]. These guidelines take into consideration the differences between 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[51] 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[41] 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[6,52]. 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 Fig. 1.

5-2 Depth images

A 360◦ RGB-D dataset similar to our conditions is Stanford- 2D-3D-S[52]. 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.

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[52] dataset are generated by projecting the labels from the 3D meshes onto the 2D images which leads to visible artifacts as seen in Fig. 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[52], PanoContext[41] and MatterportLayout[43]. Existing datasets typically assume that rooms are cuboid, which works well for standard residential spaces but fails in environments with irregular boundaries. In contrast, our dataset includes large public spaces such as offices and hospitals, where the layout often contains non-cuboid structures such as multiple connected zones. These characteristics introduce additional planes and discontinuities that challenge layout estimation models and help improve their robustness in real-world applications. Fig. 3 shows a sample of qualitative results: the MatterportLayout[43] 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.

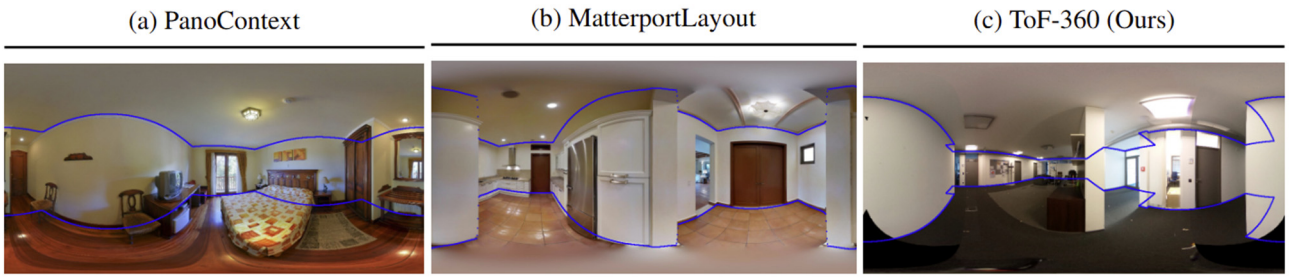

Fig. 3 Examples of annotation failure on (a) PanoContext and (b) MatterportLayout. In contrast, (c) ToF-360 provides correct room layout annotations. Ground truth boundary is shown as blue line.

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[6,13,41,52-54] 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). The current state-of-the-art approaches: HoHoNet[19], PanoFormer[21], and SFSS-MMSI[26] 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[52] and Structured3D[13] datasets and are then evaluated on our ToF-360 dataset. The Stanford-2D-3D-S[52] dataset consists of multi- modal equirectangular images with 13 object categories and divided into six areas. We use the fold_1 split for training and validation as suggested by Armeni et al.[52] The Structured3D[13] dataset is a synthetic dataset that offers 40 NYU-Depth-v2[55] 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.[13] 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[52] and 40 NYU-Depth-v2[55] object categories for Structured3D[13] 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[39] 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[39]. The model was trained on a joint dataset comprising of ScanNet[53], Structured3D[13], and S3DIS[54] datasets and validated on the S3DIS validation set. The training setup is described in detail by Wu et al.[39] and replicated for this evaluation.

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

Method	Modalities	Results		
		$mIoU(\%)$	$mAcc(\%)$	$aAcc(\%)$
HoHoNet[19]	RGB	20.76	41.65	66.90
PanoFormer[21]		28.07	51.44	**77.75**
SFSS-MMSI[26]		**29.56**	**51.53**	76.65
HoHoNet[19]	RGB-D	**27.46**	**48.66**	**76.11**
PanoFormer[21]		21.52	39.90	65.30
SFSS-MMSI[26]		24.92	45.88	73.11
SFSS-MMSI[26]	RGB-D-N	**23.17**	**46.26**	**70.39**
PTv3[39]	RGB-XYZ-N	18.57	25.08	67.89

The model achieving the best validation result was chosen for the evaluation of the ToF-360 point clouds. The S3DIS[54] dataset is the point cloud version of Stanford-2D-3D-S[52] 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. Fig. 4 and Fig. 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. 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). Fig. 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. 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 Fig. 4 and Fig. 5 we can see that the image based approaches generalized better on the furniture classes and both approaches performed similarly for the structural elements.

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 (δ1) are calculated following[42]. We used the layout estimation models provided by LGTNet[56]. These models are pre-trained by the authors with public datasets consisting of Stanford-2D-3D-S[52], PanoContext[41], and Matterport- Layout[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.

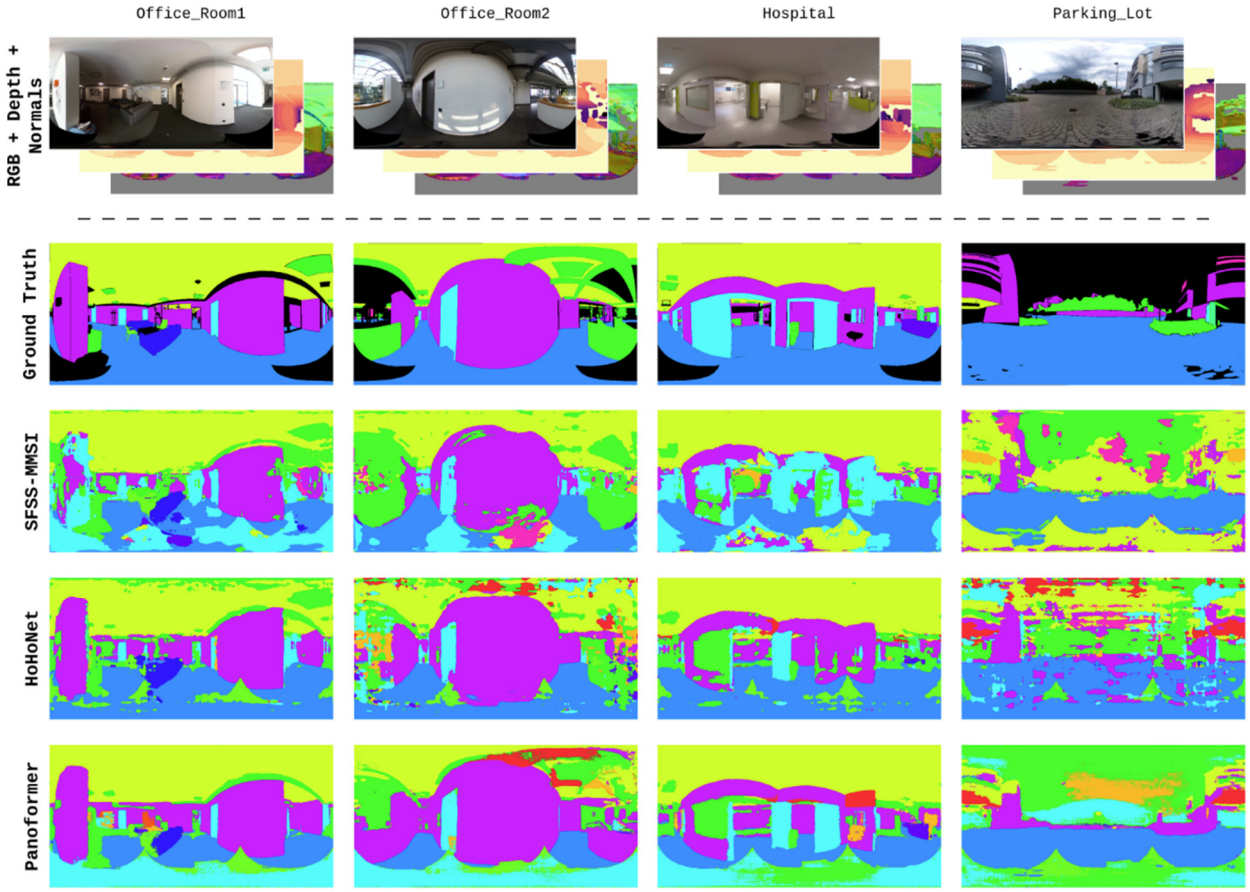

Fig. 4 Visualization of RGB-Depth-Normals semantic segmentation results for the proposed ToF-360 dataset. In the above visualization, SFSS-MMSI is trained with RGB-Depth + Normals while HoHoNet and PanoFormer with RGB-Depth panoramic equirectangular images from Stanford-2D-3D-S.

Fig. 5 Visualization of the results of the point cloud based semantic segmentation using Point Transformer V3, the colors correspond to the same classes as in Fig. 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.

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 Fig 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. IoU values are in %, for RM SE lower values are better.

Trained dataset	$2DIoU^{\uparrow}$	$3DIoU^{\uparrow}$	$RMSE^{\downarrow}$	$\delta_1^{\uparrow}$
Stanford-2D-3D-S[52]	59.66	57.33	0.742	0.831
PanoContext[41]	60.20	57.80	0.770	0.849
MatterportLayout[43]	**62.71**	**62.88**	**0.730**	**0.900**

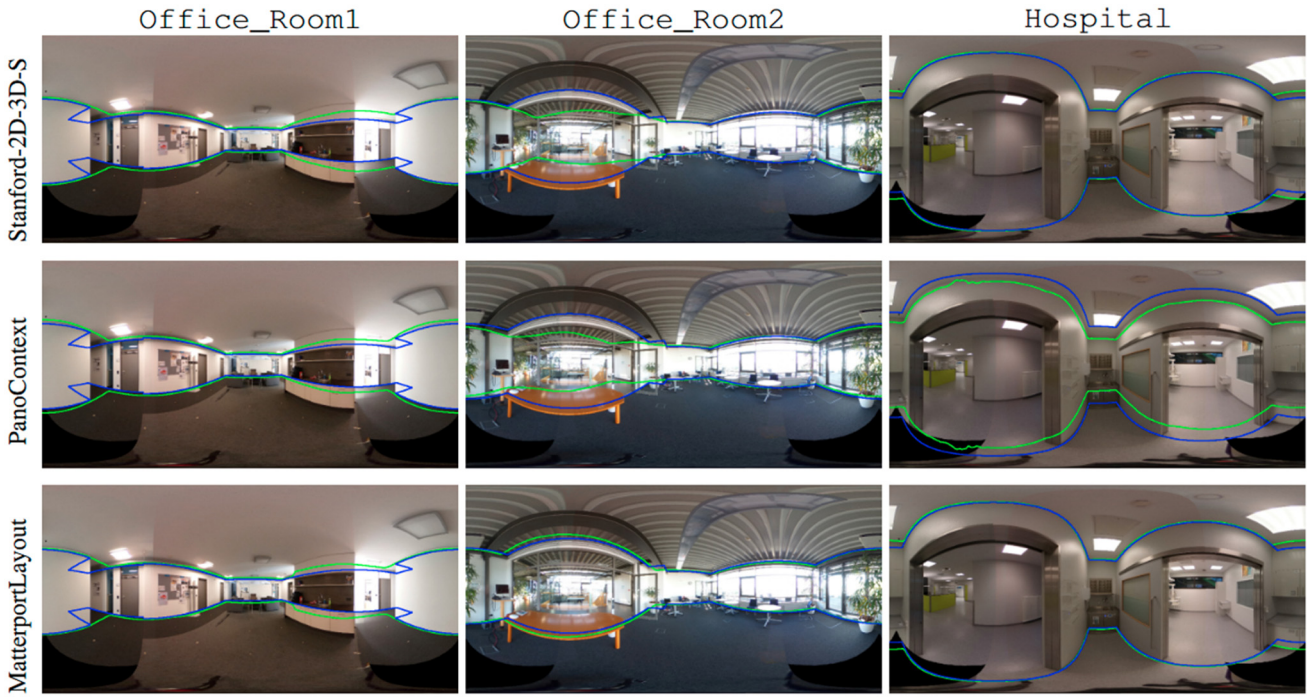

Fig. 6 Qualitative comparison of layout estimation methods on Tof-360 produced using LGT-Net. We show the boundaries of room layout on panorama. The blue lines are ground truth, and the green lines are prediction.

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, defining a new benchmark for single-shot reconstruction without the need for global alignment. Unlike existing datasets that mainly assume simple cuboid

room layouts, ToF-360 includes spaces with non-cuboid structures such as long corridors, glass partitions, and complex room connections. This makes the dataset particularly useful for evaluating how robust models trained on existing datasets are when applied to real-world scenes. As our dataset is confined to a limited number of real-world scenes, it serves as a challenge to existing works and highlights the difficulties in generalizing models trained on prior datasets, especially regarding annotation non-uniformity and scene bias. Future work will 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.

Acknowledgements

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).

References

1) W. Choi et al.: Understanding indoor scenes using 3d geometric phrases, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 33–40 (2013).
2) M. Naseer, S. Khan, F. Porikli: Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access*, 7, 1859–1887 (2018).
3) I. ALHASHIM, P. WONKA: High Quality Monocular Depth Estimation via Transfer Learning, *arXiv e-prints*, abs/1812.11941:arXiv:1812.11941 (2018).
4) Y. NIE et al.: Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
5) H. ZHAO et al.: Point Transformer, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268 (2021).
6) A. CHANG et al.: Matterport3D: Learning from RGB-D Data in Indoor Environments, *arXiv preprint*, arXiv:1709.06158 (2017).
7) Microsoft: Azure Kinect and Kinect Windows v2 comparison, Microsoft Learn, https://learn.microsoft.com/en-us/azure/kinect-dk/windows-comparison#hardware (2019).
8) Apple: iPhone 13 Pro - Technical Specifications, Apple Support, https://support.apple.com/kb/SP852?locale=en_US (2021).
9) Intel: Intel® RealSense™ LiDAR Camera L515, Product Specifications, https://ark.intel.com/content/www/us/en/ark/products/201775/intel-realsense-lidar-camera-l515.html (2019).
10) Ricoh: AI Solution for Spatial Data Creation and Utilization, Ricoh's Technology, https://www.ricoh.com/technology/tech/126_building_digital_twin (2024).
11) M. Firman: RGBD Datasets: Past, Present and Future, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 19–31, IEEE (2016).
12) S. Song et al.: Semantic Scene Completion from a Single Depth Image, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1746–1754, IEEE (2017).
13) J. Zheng et al.: Structured3D: A Large Photo-Realistic Dataset for Structured 3D Modeling, *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 519–535, Springer (2020).
14) A. Dehghan et al.: ARKitScenes: A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data, *NeurIPS Datasets and Benchmarks*, Vol. 2, No. 6, p. 16, NeurIPS (2021).

15) P. Schneider et al.: TIMo—A Dataset for Indoor Building Monitoring with a Time-of-Flight Camera, *Sensors*, Vol. 22, No. 11, p. 3992, MDPI (2022).
16) M. Eder et al.: Tangent Images for Mitigating Spherical Distortion, *CVPR*, pp. 12423–12431, Computer Vision Foundation / IEEE (2020).
17) Y. K. Lee et al.: SpherePHD: Applying CNNs on a Spherical Polyhedron Representation of 360° Images, *CVPR*, pp. 9181–9189, Computer Vision Foundation / IEEE (2019).
18) S. Orhan, Y. Bastanlar: Semantic Segmentation of Outdoor Panoramic Images, *Signal Image Video Process*, Vol. 16, No. 3, pp. 643–650 (2022).
19) C. Sun, M. Sun, H.-T. Chen: HoHoNet: 360 Indoor Holistic Understanding with Latent Horizontal Features, *IEEE/CVF CVPR* (2021).
20) Z. Zheng et al.: Complementary Bi-Directional Feature Compression for Indoor 360° Semantic Segmentation with Self-Distillation, *WACV*, pp. 4490–4499, IEEE (2023).
21) Z. Shen et al.: PanoFormer: Panorama Transformer for Indoor 360° Depth Estimation, *ECCV (1)*, pp. 195–211, Springer (2022).
22) J. Dai et al.: Deformable Convolutional Networks, *ICCV*, pp. 764–773, IEEE Computer Society (2017).
23) O. Ronneberger, P. Fischer, T. Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, *MICCAI (3)*, pp. 234–241, Springer (2015).
24) E. Xie et al.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, *NeurIPS*, pp. 12077–12090 (2021).
25) J. Guerrero-Viu et al.: What's in My Room? Object Recognition on Indoor Panoramic Images, *IEEE ICRA* (2020).
26) S. Guttikonda, J. R. Rambach: Single Frame Semantic Segmentation Using Multi-Modal Spherical Images, *WACV*, pp. 3210–3219, IEEE (2024).
27) J. Zhang et al.: Bending Reality: Distortion-Aware Transformers for Adapting to Panoramic Semantic Segmentation, *CVPR*, pp. 16896–16906, IEEE (2022).
28) J. Zhang et al.: Behind Every Domain There Is a Shift: Adapting Distortion-Aware Vision Transformers for Panoramic Semantic Segmentation, *CoRR*, abs/2207.11860 (2022).
29) X. Chen et al.: Multi-View 3D Object Detection Network for Autonomous Driving, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, IEEE (2017).
30) A. H. Lang et al.: PointPillars: Fast Encoders for Object Detection from Point Clouds, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, IEEE (2019).
31) C. Choy, J. Gwak, S. Savarese: 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3075–3084, IEEE (2019).
32) B. Graham, M. Engelcke, L. Van Der Maaten: 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9224–9232, IEEE (2018).
33) C. R. Qi et al.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, IEEE (2017).
34) C. R. Qi et al.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, *Advances in Neural Information Processing Systems*, Vol. 30 (2017).
35) Y. Wang et al.: Dynamic Graph CNN for Learning on Point Clouds, *ACM Transactions on Graphics (TOG)*, Vol. 38, No. 5, pp. 1–12 (2019).
36) A. Vaswani et al.: Attention Is All You Need, *Advances in Neural Information Processing Systems*, Vol. 30 (2017).

37) H. Zhao, J. Jia, V. Koltun: Exploring Self-Attention for Image Recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076–10085, IEEE (2020).
38) X. Wu et al.: Point Transformer v2: Grouped Vector Attention and Partition-Based Pooling, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 33330–33342 (2022).
39) X. Wu et al.: Point Transformer v3: Simpler Faster Stronger, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4840–4851, IEEE (2024).
40) C.-Y. Lee et al.: RoomNet: End-to-End Room Layout Estimation, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4865–4874, IEEE (2017).
41) Y. Zhang et al.: PanoContext: A Whole-Room 3D Context Model for Panoramic Scene Understanding, *ECCV (6)*, pp. 668–686, Springer (2014).
42) C. Zou et al.: LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image, *IEEE/CVF CVPR* (2018).
43) C. Zou et al.: Manhattan Room Layout Reconstruction from a Single 360° Image: A Comparative Study of State-of-the-Art Methods, *arXiv preprint arXiv:1910.04099* (2019).
44) Ricoh: Product Details: RICOH THETA V, RICOH THETA, https://theta360.com/en/about/theta/v.html (2023).
45) S. SONG, S. P. LICHTENBERG, J. XIAO: SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576 (2015).
46) S. Arya et al.: An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions, *JACM* (1998).
47) OpenCV: OpenCV, GitHub, https://github.com/itseez/opencv (2025).
48) X. Deng et al.: Automatic Spherical Panorama Generation with Two Fisheye Images, *2008 7th World Congress on Intelligent Control and Automation*, pp. 5955–5959, IEEE (2008).
49) Justin Brooks: COCO Annotator, GitHub, https://github.com/jsbroks/coco-annotator/ (2019).
50) F. Kaufmann et al.: Ontology-Based Semantic Labeling for RGB-D and Point Cloud Datasets, *EC3 Conference 2023*, Vol. 4, European Council on Computing in Construction (2023).
51) S.-T. Yang et al.: A Semi-Automatic Tool for Indoor Panorama Layout Annotation, *SIGGRAPH ASIA Posters*, pp. 34:1–34:2, ACM (2018).
52) I. Armeni et al.: Joint 2D-3D-Semantic Data for Indoor Scene Understanding, *arXiv preprint arXiv:1702.01105* (2017).
53) A. DAI et al.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839 (2017).
54) I. Armeni et al.: 3D Semantic Parsing of Large-Scale Indoor Spaces, *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, IEEE (2016).
55) D. Ignatov, A. Ignatov, R. Timofte: Virtually Enriched NYU Depth V2 Dataset for Monocular Depth Estimation: Do We Need Artificial Augmentation?, *CVPR*, IEEE (2024).
56) Z. Jiang et al.: LGT-Net: Indoor Panoramic Room Layout Estimation with Geometry-Aware Transformer Network, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1654–1663, IEEE (2022).