SIGNAV: Semantically-Informed GPS-Denied Navigation and Mapping in Visually-Degraded Environments

Alex Krasner*, Mikhail Sizintsev*, Abhinav Rajvanshi, Han-Pang Chiu,
Niluthpol Mithun, Kevin Kaighn, Philip Miller, Ryan Villamil, Supun Samarasekera
SRI Internaional, USA *

Abstract

Understanding the perceived scene during navigation enables intelligent robot behaviors. Current vision-based semantic SLAM (Simultaneous Localization and Mapping) systems provide these capabilities. However, their performance decreases in visually-degraded environments, that are common places for critical robotic applications, such as search and rescue missions. In this paper, we present SIGNAV, a real-time semantic SLAM system to operate in perceptually-challenging situations. To improve the robustness for navigation in dark environments, SIGNAV leverages a multi-sensor navigation architecture to fuse vision with additional sensing modalities, including an inertial measurement unit (IMU), LiDAR, and wheel odometry. A new 2.5D semantic segmentation method is also developed to combine both images and LiDAR depth maps to generate semantic labels of 3D mapped points in real time. We demonstrate that the navigation accuracy from SIGNAV in a variety of indoor environments under both normal lighting and dark conditions. SIGNAV also provides semantic scene understanding capabilities in visually-degraded environments. We also show the benefits of semantic information to SIGNAV's performance.

1. Introduction

Accurate navigation and scene understanding are key capabilities for autonomous robots to a variety of critical applications in unknown GPS-denied environments. Recent vision-based semantic SLAM (simultaneous localization and mapping) systems provide these capabilities, by performing image-based semantic segmentation techniques to assign class labels to 3D mapped points. The resulting 3D semantic map enables more intelligent robot behaviors, such as finding doors to quickly move from one room to another. It is also a more natural representation for humans to understand the mapped environment.

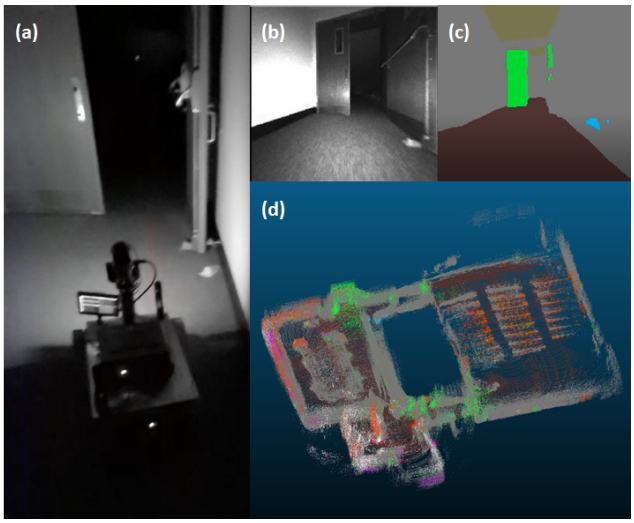

Figure 1. An example of our SIGNAV system operating in dark GPS-denied indoor environment: (a) the robot platform moves from one dark room to another, and the only lighting source is from its own LED, (b) input video frame, (c) 2.5D semantic segmentation of the input video frame, (d) a 3D semantic map produced by SIGNAV in real time. Note different colors in (c) and (d) represent different semantic classes (such as green color represents door class).

However, performance of these systems degrades dramatically in perceptually-challenging environments such as tunnels and mines, that are common places for robotic applications including infrastructure inspection [19], surveillance [15], and indoor search and rescue missions [1, 11]. Vision-based navigation methods are unreliable in dark locations even in presence of on-board lights because far scene regions are still poorly visible. The quality of image-based semantic segmentation is also poor in visually-degraded situations. While LiDAR-based SLAM algorithms are more robust to illumination variations, they are erroneous in places with geometrically self-similar patterns such as long corridors.

*The first two authors contributed equally to this work. All authors are with Center for Vision Technologies, SRI International, USA. The contact author is Han-Pang Chiu {han-pang.chiu@sri.com).

In this paper, we present SIGNAV (Semantically-Informed Gps-denied NAVigation), a real-time semantic SLAM system that enables robust navigation and scene understanding capabilities in visually-degraded environments (Figure 1). To address navigation challenges in dark locations, SIGNAV leverages a multi-sensor navigation architecture [5] to fuse vision with additional sensing modalities, including an inertial measurement unit (IMU), LiDAR, and wheel odometry. This approach combines the strengths of different sensors to improve navigation accuracy. It also provides more robust estimation than using a single sensor, by avoiding the single point of failure in navigation.

To generate a reasonable 3D semantic map in perceptually-challenging situations, SIGNAV utilizes a novel 2.5D semantic segmentation method that combines both gray-scale camera images and LiDAR depth maps for semantic labeling of 3D mapped points. SIGNAV also uses the semantic labels to identify and remove features (also related 3D mapped points) from non-rigid classes (e.g. people) and non-Lambertian surfaces. Focusing on static and rigid classes improves the quality of loop detection process.

We verified and demonstrated SIGNAV using a ground vehicle (GVR-bot) with a variety of scenarios in GPS-denied indoor environments, under both normal lighting and dark conditions. We compare SIGNAV's performance with several state-of-the-art SLAM methods using the data acquired from these test scenarios. We also showed the benefits of semantic information from SIGNAV in visually-degraded environments.

The rest of the paper is organized as follows. In Section II, we present the related work and highlight our contributions. In Section III, we describe SIGNAV, including the details of each system module. In Section IV, we present our experimental setup and results for different scenarios. Conclusions and future work are presented in Section V.

2. Related Work

In this section, we provide a brief review on SLAM systems in perceptually-challenging environments. We refer the readers to [3] for a broad survey on SLAM.

2.1. Navigation in Visually-Degraded Environments

Relying on single sensor modality is challenging for navigation in visually-degraded environments. For example, vision-based SLAM methods [3] work well under normal lighting conditions. However, their performance degrades in dark environments due to poor quality of image data.

Among all sensor choices, LiDAR sensors provide long-range 3D measurements without external lighting sources. Therefore, 3D LiDAR SLAM systems [10, 34, 16, 20, 45, 23, 37, 28, 42] have been popular for autonomous robots in GPS-denied dark locations, such as subterranean environments. However, LiDAR-based systems tend to fail in environments with geometrically self-similar patterns such as long hallways. The recent trend is to fuse LiDAR sensors with different sensor modalities, such as inertial sensors [30, 35, 41] or visual-inertial odometry [43, 36], to improve its performance in perceptually-challenging environments.

Our work follows this trend to combine measurements from a set of sensors to improve the quality of the solution in perceptually-challenging environments. In addition, our system is designed to provide scene understanding capabilities under dark conditions.

2.2. Semantic SLAM

There has been a surge of interest towards real-time semantic SLAM systems in recent years. Most of these works [24, 40, 21, 33, 32, 39, 31, 25, 14, 22, 44, 9, 27] rely on RGBD cameras or monocular cameras. The typical approach is to utilize pre-trained deep networks to assign semantic labels to each imaged pixel on an input video frame in real time. It then associates these semantic labels between 2D imaged pixels and correspondent 3D mapped points. However, 2D semantic segmentation quality from these systems significantly decreases in visually-degraded environments.

Recently, there are also LiDAR-based semantic SLAM methods [2, 12] that use pre-trained deep networks to generate semantic labels on 2D range maps, which are projected from high-quality 3D LiDAR point clouds. These methods utilize data from large and expensive LiDAR units for outdoor self-driving car applications. In contrast, our system is designed with low-cost sensors[1] for smaller autonomous robots operating in indoor and subterranean environments (such as mines). Since the quality (density, coverage, and measurement accuracy) of our LiDAR sensor is much lower, we need to fuse additional sensor modalities to improve both SLAM accuracy and semantic labeling quality in perceptually-challenging environments.

2.3. Contribution

To the best of our knowledge, SIGNAV is the first real-time semantic SLAM system designed for operating in visually-degraded indoor and subterranean environments. The highlights of SIGNAV are as follows:

- **Multi-Sensor SLAM:** SIGNAV starts with a tightly-coupled visual-inertial SLAM system, and integrates LiDAR odometry measurements and wheel odometry readings in a loosely-coupled manner. It leverages a flexible plug-and-play architecture [5] based on factor graphs for multi-sensor navigation. The combination

[1] LiDAR-based semantic SLAM systems use velodyne hdl-64e LiDAR, which costs around 75,000 dollars. In contrast, our total sensor cost (inertial, camera, and two low-cost LiDAR units) is less than 6,500 dollars.

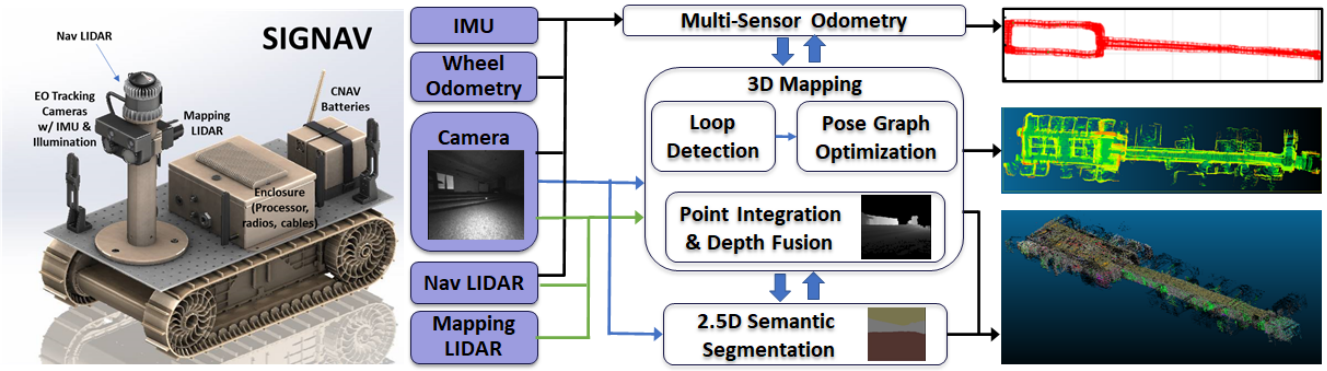

Figure 2. SIGNAV's system pipeline: three major system modules - multi-sensor odometry (Section 3.2), 3D Mapping (Section 3.3), and 2.5D semantic segmentation (Section 3.4) - that process sensor data (IMU, stereo cameras, LiDAR, wheel odometry) to generate real-time output (from top to bottom: accumulated pose estimates, 3D mapped points, a 3D semantic map).

of these sensing modalities improves the robustness and accuracy for navigation in dark environments.

- **2.5D Semantic Segmentation:** SIGNAV incorporates a new 2.5D semantic segmentation technique, which combines both gray-scale monocular images and LiDAR depth maps, to generate reasonable real-time semantic labels in dark environments. This 2.5D semantic segmentation technique utilizes a mixture-of-expert architecture (UNO [38]) to fuse results from two pre-trained deep learning networks: one network processes images, and the other network handles depth maps. The entire computation of SIGNAV (including 2.5D semantic segmentation) is enabled using one embedded processor unit (NVIDIA Xavier).

- **Semantically-Informed Mapping:** SIGNAV refines the 3D mapped points by identifying and removing non-rigid classes (objects and people that are not part of the static scene) as well as non-Lambertian surface classes (since these areas are particularly challenging for computer vision processing). This allows the SLAM process to use only features and mapped points correspondent to static and rigid object classes. Such an approach improves (1) the quality of loop detection and (2) the pose estimation accuracy when matching the new image to the 3D map.

3. SIGNAV

In this section, we describe SIGNAV's system pipeline (Figure 2) with details of each system module.

3.1. Sensing Modalities

SIGNAV is a multi-sensor semantic SLAM system designed for operating in perceptually-challenging indoor and subterranean environments. We utilize a sensor fusion framework [5] based on factor graphs, which is capable of incorporating multiple sensors with different rates, latencies, and error characteristics. Factor graphs have been used [8] for many applications related to robotic navigation. They naturally encode the factored nature of the probability density over the navigation states (3D position, 3D orientation, and 3D velocity at any given time in our case), clearly separating the state representation from the constraints induced by the sensor measurements. The connectivity of the factor graph defines which state variables are affected by which sensor measurements. This representation makes it ideal for fusing multiple sensors for navigation.

SIGNAV currently integrates sensor measurements from four sensor modalities (IMU, cameras, LiDAR, and wheel odometry) using this plug-and-play factor graph framework. Note for current hardware, we use two LiDAR units: the forward-facing navigation 3D LiDAR contributes to both multi-sensor odometry and 3D mapping, while the upward-facing mapping 2D LiDAR is only used to increase the coverage of 3D mapping of the perceived environment.

3.2. Multi-Sensor Odometry

SIGNAV utilizes a parallel architecture to simultaneously compute the robot's motion over time (multi-sensor odometry) and model its perceived environment (3D mapping and 2.5D semantic segmentation). SIGNAV's multi-sensor odometry module is essentially a multi-sensor fusion process that starts with a tightly-coupled visual-inertial odometry mechanism [13] to fuse IMU (Inertial Measurement Units) data and camera feature track measurements. Inertial measurements from IMU are produced at a much higher rate than other sensors. Therefore, we summarize multiple consecutive inertial readings between two navigation states created at the time when other sensor measurements come (such as camera features from a video frame). This IMU factor generates 6 degrees of freedom relative pose and corresponding velocity change as the underlying

motion model, that replaces traditional process models in vision-based SLAM systems.

SIGNAV further integrates sensor measurements from LiDAR and wheel odometry in a loosly-coupled manner. For LiDAR, we use the Fast-GICP algorithm [17] to perform efficient voxel-based generalized ICP (Iterative Closet Point) process to register 3D LiDAR points obtained from sequential scans (scan-to-scan registration). A 3D related pose measurement across sequential scans is then generated and fused within our factor graph architecture.

SIGNAV integrates wheel odometry readings inside the multi-sensor odometry module as 3D velocity measurements, rather than relative pose constraints or simple speed. While it is also capable to capture the rotation reading from differential wheel motion, we found this information is unreliable from our experiments resulting in noticeably curving trajectories. Therefore, rather than using a speed scalar constraint only, we formulate wheel odometry factor as a 3D velocity vector that constraints speed in a local direction of the vehicle that naturally encompasses backward and forward motions. We found this explicit velocity formulation from wheel odometry is beneficial to typical visual-inertial SLAM systems when used on our wheeled and tracked robots. It improves the navigation accuracy when robot locomotion has high-frequency vibrations, that reduce reliability of the IMU estimates and affects the overall scale estimation of the reconstructed scene. Another important capability is to provide zero-velocity update (e.g. when robot is stationary), that avoids potential drift especially in dark environments where the amount of valid visual features decreases and temporal feature track length is short.

3.3. 3D Mapping

The 3D mapping module in SIGNAV is enabled by loop detection and pose graph optimization. Loop detection sub-module establishes associations (loops) across non-consecutive video frames taken at different times (when a robot revisits the same place). These associations are used to optimize the past poses involved within the loops. Both multi-sensor odometry poses and loop-closure optimized poses are used to continuously integrate 3D mapped LiDAR points accumulated from past scans during the run.

3.3.1 Loop Detection and Pose Graph Optimization

During navigation, SIGNAV selects key frames from input video streams, and adds them into the database. Note that database entry is essentially a video frame that holds the collection of keypoints with their descriptors, image locations, and 3D world coordinates computed from triangulation across matched stereo 2D points across video frames. The selection is based on conditions between new frame and past key frames, including the number of overlapped features, the temporal difference, and the spatial difference between poses associated with frames.

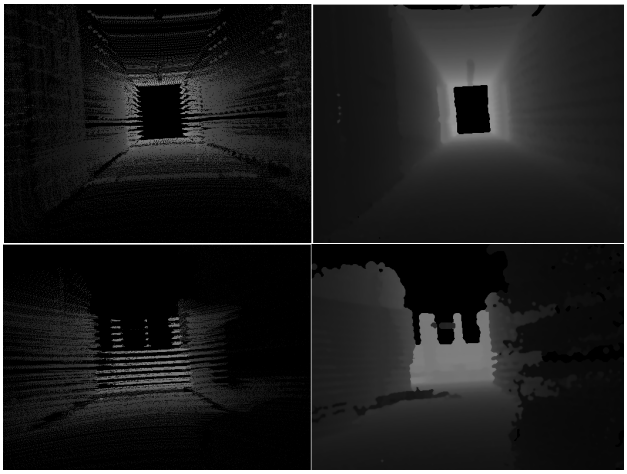

Figure 3. Examples of LiDAR depth maps generated before (left) and after (right) the depth completion algorithm [18].

Loop detection is achieved by matching new image to the database of key frames. If a frame is matched to a keyframe that has been added before, it indicates the matched keyframe is acquired when the vehicle previously visited the same place. Therefore, these matches can be treated as loop closures to optimize past poses involved within the loop, which is the typical pose graph optimization process. The optimized pose is then fed back to the multi-sensor odometry module to correct the drift for real-time navigation solution. This process is adapted from [6].

3.3.2 Point Integration and Depth Fusion

We use both real-time estimated pose solution (from multi-sensor odometry module) and optimized poses (from pose graph optimization) to iteratively re-integrate involved 3D LiDAR points from past scans. First, appropriate LiDAR poses are interpolated from the pose solution, since the frequencies may not be the same. Individual LiDAR scans are then transformed to these interpolated poses. Overlapping points between consecutive LiDAR scans are removed. Thus, this process continuously produces a 3D map of point clouds during navigation.

The transformed points are then passed to the depth fusion process. This process leverages the depth and range information present in the LiDAR point cloud to better enable 2.5D semantic understanding of the scene, especially in dark environments where 2D image quality is poor.

The low-cost Ouster OS1-16 LiDAR we use in our robot only yields 16 vertical beams during scanning resulting in a low vertical resolution 3D point cloud. To mitigate this issue, we first temporally aggregate LiDAR scans over a past 1-second window. This way produces a denser Li-

DAR point cloud at frame t-1 (seconds). The aggregated point cloud is then projected to the camera image plane producing a depth map that can be used for segmentation. The 1-second sliding window was determined experimentally as the one producing adequate aggregated point clouds at robot locomotion speeds without significant delay to the navigation algorithm. Despite the temporal integration of LiDAR scans, the projected depth image is often too sparse for subsequent out-of-the box segmentation algorithms that assume input depth images with all pixels defined. To overcome this sparsity limitation, we use the IP-Basic depth completion algorithm [18] to produce more convincing and consistent dense depth maps that are used as subsequent input to 2.5D semantic segmentation module. Figure 3 depicts two examples of 1-second integrated LiDAR point clouds and corresponding depth completed results.

3.4. 2.5D Semantic Segmentation

SIGNAV introduces a novel 2.5D semantic segmentation method to combine both 2D image and LiDAR depth maps for semantic labeling of 3D point clouds in visually-degraded environments. Note there exist powerful deep neural network models which can generate robust semantic labels for RGB images. However, these segmentation networks are highly sensitive to lighting conditions and can fail catastrophically in the visually degraded and dark environments. Semantic segmentation directly on 3D LiDAR point cloud can help in these kinds of environments. However, they typically require extensive computational resource and cannot provide real time inference on small SWAP (size, weight, and power) machines.

In this regard, we present a 2.5D semantic segmentation method to SIGNAV, which is robust to varying lighting conditions and can run using GPUs from an embedded processor. We use this 2.5D semantic segmentation module to label selected video frames (key frames) with correspondent depth maps (from point integration and depth fusion submodule), for real-time 3D semantic mapping during navigation. Our method has two networks operating independently on gray-scale image and depth data. For the two networks, we choose a pre-trained DeepLab based network [4] (originally for RGB semantic segmentation) trained with the backbone of Xception65 [7] on ADE20K dataset.

For our image segmentation network, we converted ADE20K dataset to gray-scale and fine-tuned the model with those images. However, for our depth semantic segmentation network, we do not have this kind of pre-trained network that generalizes reasonably to our target environments. Therefore, we train the depth segmentation network using a weakly-supervised learning approach based on cross-modal supervision inspired from [26], as visualized in Figure 4. Specifically, we collected RGB+Depth (RGBD) data in various well-lit environments to get the

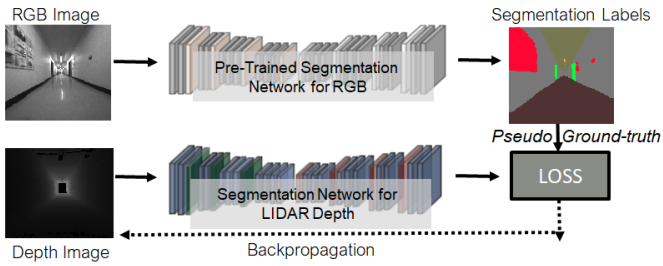

Figure 4. The concept of our weakly-supervised approach to train depth segmentation network using semantic outputs (as pseudo ground truth) from pre-trained image segmentation network on paired RGB and Depth data.

best-possible RGB segmentation output. Then, the output of the pre-trained RGB segmentation network is considered as a pseudo- ground truth for the depth segmentation network (ideally, these two network should produce the same segmentation results). In turn, this pseudo-ground truth is used to provide weak supervision for training the depth segmentation network. Note this supervision only happens in the training stage. The trained depth segmentation network can generate its own semantic segmentation results without any information from the image segmentation network.

During navigation, the 2.5D semantic segmentation module fuses the semantic segmentation results from these two trained networks (grayscale and depth) using a fusion algorithm adapted from UNO[38]. UNO is an uncertainty-aware fusion scheme to effectively fuse inputs that might suffer from a range of known and unknown degradation and compensate for errors caused by out-of-distribution conditions. At training time, we compute the entropy of each network over our training dataset by tracking the class label probabilities of each network. This establishes the baseline performance of each network, and determines whether inference results lie within the expected distribution. At inference-time (during navigation), each network produces the probabilities for each label (per pixel) – then these probabilities are re-balanced and re-scaled based on how likely they are to be the part of the training distribution, that tracks whether the new input is too degraded for a network to accurately make a prediction. Finally, these re-balanced outputs of each network are combined to produce the final semantic label probabilities for each pixel using a Noisy-Or operation as follows.

$$I(y = c) = 1 - \Pi_i(1 - p_i(y = c|x_i, \theta_i)) \quad (1)$$

$$p(y = c) = \frac{I_c}{\Sigma_j I_j} \quad j = 1, 2, ..., C \quad (2)$$

where $p_i(y = c|x_i, \theta_i)$ is the predictive probability of network i (image or depth) for class c (total C classes), x_i and θ_i are the input and parameters of network i and p_c is the final probability for class c.

The semantically labeled image is then back-projected onto its parent 3D map to label individual points of the LiDAR point cloud. The annotated 3D points are then accumulated to produce a semantically labeled 3D map. SIGNAV also refines the 3D mapped points based on the 2.5D semantic segmentation results, by identifying and removing non-rigid classes (objects that are not part of the static scene) and non-Lambertian surfaces. Only mapped points belonging to static and rigid classes will be used in subsequent visual matching for loop closure detection.

4. Experimental Evaluation

In this section, we present the experimental results obtained from tests in a variety of GPS-denied indoor environments under normal lighting or dark conditions. We also collected the data from four test scenarios to compare SIGNAV with state-of-the-art SLAM solutions, and perform an analysis on SIGNAV capabilities.

4.1. Experimental Setup

Each dataset includes data from four sensor modalities: Electro-Optical (EO) stereo cameras, inertial sensing, LiDAR, and wheel odometry. All sensors are installed on a GVR-bot platform (Figure 2) for experiments and data collection. The GVR-bot platform has fast moving speed - around 1 meter per second for our experiments. Wheel odometry is directly retrieved from ROS topics provided by the GVR-bot. We used an Intel Realsense T-265 unit, which is equipped with fisheye 170 degree field-of-view gray-scale stereo camera and an IMU. We operated the camera with 848x800 resolution at global shutter, and recorded video data at 15Hz. The IMU is recorded at 200Hz. The camera and IMU readings are synchronized with the exact time offset being determined during the initial calibration process. The camera is mounted forward-facing, and also equipped with 2 LED lights to help navigating in the dark environments.

We installed two low-cost LiDAR sensors[1] , which are more affordable for indoor robots. The navigation LiDAR (Ouster OS1 3D LiDAR) is mounted as a standard horizontal setup to contribute to both navigation and 3D mapping. We operated and recorded it at 5Hz. The mapping LiDAR (Hokuyo UST-20LX 2D LiDAR) is recorded at 20Hz frequency. This sensor is oriented vertically perpendicular to the optical axis of the front-facing camera, (a.k.a coronal) in order to capture slices of ceiling, walls and floor during robot navigation. This setup increases the coverage for 3D mapping of the perceived environment. However, this mapping LiDAR does not contribute to navigation estimation, since scans always cover different portion of surfaces while robot is moving.

We conducted each of four scenarios for around 5 minutes at its respective indoor environments. The first two scenarios are under normal lighting conditions. Scenario 1 is for the robot to operate two full loops inside a building including long hallways, while Scenario 2 is inside a cubicle office environment to conduct three loops with different route variations (some portions of the route are repeated, some are not). We set up perceptually-challenging environments for the final two scenarios. Scenario 3 is to navigate across two conference rooms (conduct two loops - with partially different routes - in each room): one room is totally dark, while the other room has little light from outside. Scenario 4 is to operate inside a dark auditorium with three repeated loops. Note for both scenarios, we started and ended at an entrance with normal lighting, and turned off all external lighting sources in the navigated rooms and auditorium. All SIGNAV computation (including 2.5D semantic segmentation on GPUs) are enabled using one embedded processor (Nvidia Xavier) on the GVR-bot in real-time. The 2.5D semantic segmentation module runs at 1.7Hz to label the selected video frames (key frames) during navigation.

We marked a set of surveyed points (as ground truth) along the path on the ground in each test environment. The positions of these marked points are measured using state-of-the-art indoor land surveying techniques from civil engineering industry. Note we measured our platform height beforehand, so we can compute 3D navigation error using surveyed points on the ground.

4.2. Evaluation

We compare SIGNAV with several state-of-the-art 3D SLAM systems, which are used for large-scale GPS-denied navigation applications, on these datasets. LeGO-LOAM [34] is a state-of-the-art LiDAR-based SLAM system. LIO-SAM [35] augments LeoGO-LOAM with IMU measurements in a smoothing-and-mapping framework, and has recently demonstrated impressive results in visually-degraded environments. Cam-SLAM [29] is a representative tightly-coupled visual-inertial SLAM system for large-scale applications. There are also SLAM systems [43, 36], which fuse both visual-inertial odometry and LiDAR measurements. However, these systems do not have open source implementations available for us to conduct the comparison.

Each system is evaluated on all four datasets, with only input sensor adjustment. Wheel odometry is the odometry input, if used. Both LeGO-LOAM and LIO-SAM use the navigation LiDAR (Ouster OS1 3D LiDAR) and the IMU, while Cam-SLAM incorporates the stereo cameras and IMU. SIGNAV can use two LiDAR units, but only the front-facing navigation LiDAR (Ouster OS1 3D LiDAR) contributes to the navigation estimation. Loop detection capabilities are enabled in all systems.

We have also tried Kimera [31], a representative open-source semantic SLAM system. However, the estimated navigation trajectory diverges significantly, even under nor-

Table 1. The comparison of navigation accuracy (absolute position error in meters) on four datasets.

	(1) Long Hallway			(2) Cubicle Office			(3) Conference Rooms			(4) Large Auditorium		
	mean	max	std	mean	max	std	mean	max	std	mean	max	std
SIGNAV	**0.5946**	**1.6592**	**0.3970**	**0.3436**	**0.9735**	**0.2528**	**0.3004**	**0.9365**	**0.1918**	**0.2647**	1.3068	0.2562
CamSLAM	0.8254	2.5700	0.6281	**0.2751**	1.0863	**0.2406**	1.6998	4.7151	0.9204	2.3373	9.3099	2.3741
LeGO-LOAM	0.9599	3.0776	0.9155	0.8209	2.6214	0.9078	1.2785	5.9764	1.5722	**0.2357**	**1.0017**	**0.2137**
LIO-SAM	**0.6214**	**1.7039**	**0.3876**	0.3978	1.5559	0.2733	**0.3166**	**0.9211**	**0.1974**	0.4159	**0.8361**	**0.2251**
SIGNAV (no LiDAR)	0.7487	2.8474	0.7170	0.3791	**0.9454**	0.2802	0.3401	1.2092	0.2487	0.3368	1.4336	0.2902

*Top two methods for each metric are highlighted in blue and green colors.

mal lighting situations, on our datasets. It is possibly that many parameters (such as sensor noise model) have to be re-tuned for our used cameras and inertial sensors. Semantic segmentation from Kimera does not work for images under dark conditions either. Therefore, we omit Kimera from our evaluation.

Table 1 summarizes the navigation accuracy for all evaluated methods on the four datasets. The results show that SIGNAV is comparable (top two in most metrics) to state-of-the-art SLAM systems, with best results at long hallways (normal lighting) and conference rooms (dark conditions). The geometrically self-similar patterns in long hallways (Scenario 1) degrades the performance of LeGO-LOAM. LIO-SAM incorporates IMU measurements in the tightly-coupled manner to improve accuracy for this situation. CamSLAM also exhibits inferior performance due to lower number of camera feature tracks in the presence of texturless walls and shiny floor. SIGNAV achieves best accuracy, by leveraging the combination of visual-inertial odometry and LiDAR measurements.

With lots of textures in the cubicle office space (Scenario 2) under normal light conditions, both CamSLAM and SIGNAV leverage vision-based loop detection capabilities to optimize multiple repeated loops to achieve better results than LiDAR-based SLAM systems.

LIO-SAM performs well in scenarios under dark conditions (Scenario 3 and Scenario 4), while the overall performance of CamSLAM decreases dramatically as expected in these visually-degraded environments even if we use onboard lights. Specifically, CamSLAM becomes prone to occasional but very noticeable trajectory drifts when number of tracked camera features becomes critically low in dark locations. LiDAR-based systems are better suited for loop detection in dark environments - both LIO-SAM and LeGO-LOAM performs well in Scenario 4 (multiple loops in dark auditorium). SIGNAV combines strengths from multiple sensing modalities to achieve comparable accuracy to LIO-SAM in Scenario 3 under visually-degraded conditions. However, the reliance on camera-only loop detection limits its performance in Scenario 4.

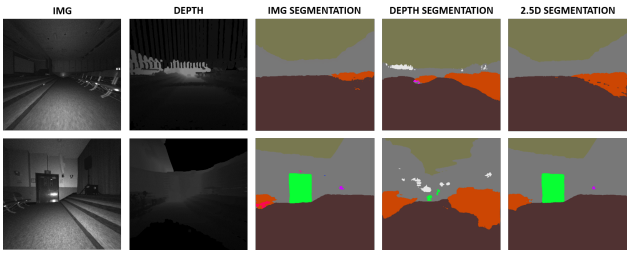

Figure 5. Two examples of 2.5D semantic segmentation from SIGNAV in darkness: (from left to right): input video frame, fused depth map, image segmentation, depth segmentation, and fused 2.5D segmentation. Note the noise from segmentation using single modality is removed in fused 2.5D semantic segmentation. Different colors represent different semantic classes. The palette for semantic segmentation includes: Ceiling, Wall, Floor, Door, Chair, Window.

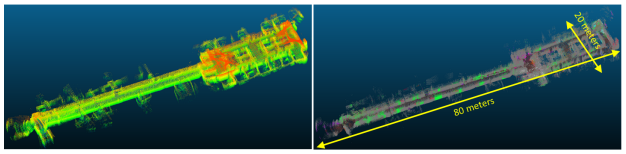

Figure 6. The 3D maps generated from SIGNAV in Scenario 1: (left) the 3D map, and (right) the 3D semantic map. Different colors in the 3D semantic map represent different semantic classes.

4.3. Impact from LiDAR and Semantic Information

We further analyze the influence of LiDAR fusion from SIGNAV to navigation accuracy in Table 1. SIGNAV (no LiDAR) demonstrates SIGNAV accuracy without LiDAR sensors. It clearly shows that the incorporation of LiDAR measurements (SIGNAV) improves accuracy for all scenarios, compared to its visual-inertial SLAM version. Note even without LiDAR sensors, the performance of SIGNAV (no LiDAR) is still comparable to other state-of-the-art SLAM systems in Table 1.

Here we also show the impact of 2.5D semantic segmentation from SIGNAV. Note the evaluation of semantic segmentation accuracy with RGBD data in visually-degraded environment is challenging, since ground truth (semantic labels) in such situations is difficult to be obtained. However, from visualization (Figure 5), we can see the fusion of both images and depth maps clearly improves the seman-

Table 2. The influence of 2.5D semantic segmentation to SIGNAV's navigation performance on four datasets: quality (the average percentage of inlier feature points from loop detection) and accuracy (the mean absolute position error in meters for the overall scenario).

	(1) Long Hallway		(2) Cubicle Office		(3) Conference Rooms		(4) Auditorium	
	quality	accuracy	quality	accuracy	quality	accuracy	quality	accuracy
SIGNAV	59.1%	0.5946	59.8%	0.3436	51.5%	0.3004	52.0%	0.2647
SIGNAV (no segmentation)	55.2%	0.7487	57.2%	0.3791	46.0%	0.3401	46.0%	0.3368

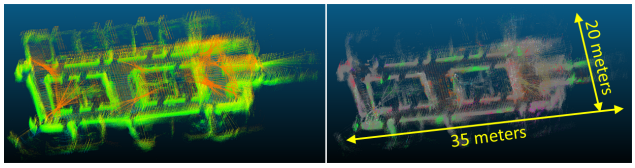

Figure 7. The 3D maps generated from SIGNAV in Scenario 2: (left) the 3D map, and (right) the 3D semantic map. Different colors in the 3D semantic map represent different semantic classes.

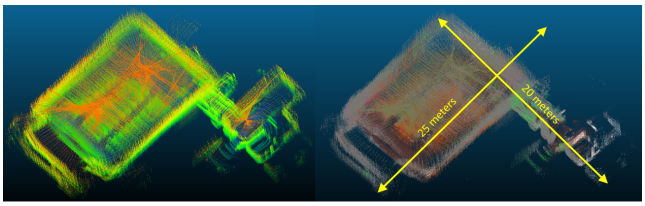

Figure 8. The 3D maps generated from SIGNAV in Scenario 4: (left) the 3D map, and (right) the 3D semantic map. Different colors in the 3D semantic map represent different semantic classes.

tic segmentation quality (noise removal), comparing to results using only one modality (image or depth), in visually-degraded environments.

We also show the 3D maps (both with and without semantic labels) generated from SIGNAV under normal lighting (Figure 6 and Figure 7) and inside dark environments (Figure 8). Note that the 2D LiDAR we used is upward facing, and it maps the structure such as ceiling and floor. Therefore, for better visualization inside the environment, we reduce the sampled 3D LiDAR points from both LiDARs. For Figure 6, there are many offices (green color - office doors, some doors are open) on both sides along the long hall way. For Figure 7, the cubicle arrangement can be seem in the middle. There are also personal offices (outside of the middle cubicle) on the top side, left side, and the bottom side. If the office door is open, SIGNAV also maps the inside structure. We can see the doors for all eight offices on the top side are open. In Figure 8, we remove the 2D LiDAR in visualization to show more details inside the auditorium. There are many chairs (red color) inside the auditorium, and the structure for those chairs are preserved.

The use of 2.5D semantic segmentation also makes the loop detection process more accurate. As shown in table 2, it improves loop detection quality, by focusing only static and rigid classes from the 3D semantic map during navigation. For all four scenarios, the percentage of inlier feature points from loop detection increases. The navigation accuracy is also slightly improved (position error decreases), because the quality of loop detection (features from non-rigid classes and non-Lambertian surfaces are removed) is increased. Most of our scenarios is conducted in static environments. We expect this approach can further improve navigation accuracy in more dynamic environments.

5. Conclusions

We present a new real-time semantic SLAM system, SIGNAV, that provides robust navigation and scene understanding capabilities within a variety of GPS-denied indoor environments, including dark places. SIGNAV incorporates LiDAR odometry and wheel odometry measurements on top of a tightly-coupled visual-inertial SLAM system. It also utilizes a new 2.5D semantic segmentation technique to combine both gray-scale monocular images and LiDAR depth maps, to generate reasonable real-time semantic labels in dark environments. SIGNAV also refines 3D mapped points based on semantic labels to improve the loop detection quality. The entire SIGNAV computation is enabled using an embedded processor unit (Nvidia Xavier).

We show the navigation accuracy of SIGNAV is comparable (or better) to other state-of-the-art SLAM systems under both normal lighting situations and visually-degraded environments. We also show the improvements from LiDAR fusion and semantic information to our performance.

Future work is to tightly integrate LiDAR measurements (instead of loosely-coupled fusion) to further improve navigation accuracy. Combining LiDAR and vision for loop detection shall also improve the performance. We expect both extensions will make SIGNAV more robust to perceptually-challenging situations and dynamic environments.

Acknowledgments

This material is based upon work supported by the Collaborative GPS-Denied Navigation for Combat Vehicles Program under Contract W9132V19C0003. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US government. We would like to thank Garry P. Glaspell, Delman B. Delbosque, and Jean D. Nelson for their valuable feedback from the project.

References

[1] H. Balta, J. Bedkowski, S. Govindaraj, K. Majek, P. Musialik, D. Serrano, K. Alexis, R. Siegwart, and G. Cubber. Integrated data management for a fleet of search-and-rescue robots. *Journal of Field Robotics*, 2016.

[2] J. Behley and C. Stachniss. Efficient surfel-based SLAM using 3D laser range data in urban environments. In *Robotics: Science and Systems (RSS)*, 2018.

[3] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Saramuzza, J. Neira, I. Reid, and J. Leonard. Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age. 32(6):1309–1332, 2016.

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[5] H. Chiu, X. Zhou, L. Carlone, F. Dellaert, and S. Samarasekera. Constrained optimal selection for multi-sensor robot navigation using plug-and-play factor graphs. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 663–670. IEEE, 2014.

[6] Han-Pang Chiu, Stephen Williams, Frank Dellaert, Supun Samarasekera, and Rakesh Kumar. Robust vision-aided navigation using sliding-window factor graphs. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 46–53. IEEE, 2013.

[7] F. Chollet. Deep learning with depthwise separable convolutions. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[8] F. Dellaert and M. Kaess. Factor graphs for robot perception. *Foundations and Trends in Robotics*, 6(1):1–139, 2017.

[9] J. Dong, X. Fei, and S. Soatto. Visual-inertial-semantic scene representation for 3D object detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[10] K. Ebadi, Y. Chang, M. Palieri, A. Stephens, A. Hatteland, E. Heiden, A. Thakur, N. Funabiki, B. Morrell, S. Wood, L. Carlone, and A. Agha-mohammadi. LAMP: Large-scale autonomous mapping and positioning for exploration of perceptually-degraded subterranean environments. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[11] T. Tomic et al. Toward a fully autonomous uav: Research platform for indoor and outdoor urban search and rescue. *IEEE robotics and automation magazine*, 19(3):46–56, 2012.

[12] X. Chen et al. SuMa++: Efficient lidar-based semantic SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

[13] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In *Robotics: Science and Systems (RSS)*, 2015.

[14] M. Grinvald, F. Furrer, T. Novkovic, J. Chung, C. Cadena, R. Siegwart, and J. Nieto. Volumetric instance-aware semantic mapping and 3D object discovery. 4(3):3037–3044, 2019.

[15] B. Grocholsky, J. Keller, V. Kumar, and G. Pappas. Cooperative air and ground surveillance. *IEEE robotics and automation magazine*, 13(3):16–25, 2006.

[16] X. Ji, L. Zuo, C. Zhang, and Y. Liu. LLOAM: LiDAR odometry and mapping with loop-closure detection based correction. In *IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2019.

[17] K. Koide, M. Yokozuka, S. Oishi, and A. Banno. Voxelized gisp for fast and accurate 3D point cloud registration. *EasyChair Preprint*, 2703, 2020.

[18] J. Ku, A. Harakeh, and S. Waslander. In defense of classical image processing: Fast depth completion on the CPU. In *International Conference on Computer and Robot Vision (CRV)*, 2018.

[19] D. Lattanzi and G. Miller. Review of robotic infrastructure inspection systems. *Journal of Infrastructure Systems*, 23(3), 2017.

[20] M. Leingartner, J. Maurer, A. Ferrein, and G. Steinbauer. Evaluation of sensors and mapping approaches for disaters in tunnels. *Journal of field robotics*, 33(8):1037–1057, 2016.

[21] C. Li, H. Xiao, K. Tateno, F. Tombari, N. Navab, and G. Hager. Incremental scene understanding on dense SLAM. In *IEEE/RSJ International Conference on Robots and Systems (IROS)*, 2016.

[22] K. Lianos, J. Schonberger, M. Pollefeys, and T. Sattler. Vso: Visual semantic odometry. In *European Conference on Computer Vision (ECCV)*. Springer, 2018.

[23] F. Mascarich, S. Khattak, C. Papachristos, and K. Alexis. A multi-modal mapping unit for autonomous exploration and mapping of underground tunnels. In *IEEE Conference on Aerospace*. IEEE, 2018.

[24] J. McCormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger. Fusion++: Volumetric object-level SLAM. In *International Conference on 3D Vision (3DV)*, 2018.

[25] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. Semanticfusion: Dense 3D semantic mapping with convolutional neural networks. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

[26] N. Mithun, K. Sikka, H. Chiu, S. Samarasekera, and R. Kumar. RGB2LIDAR: Towards solving large-scale cross-modal visual localization. In *ACM Multimedia Conference (MM)*. ACM, 2020.

[27] G. Narita, T. Seno, T. Ishikawa, and Y. Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. *arXiv preprint arXiv:1903.01177*, 2019.

[28] A. Nuchter, H. Surmann, K. Lingemann, J. Hertzberg, and S. Thurn. 6D SLAM with an application in autonomous mine mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2004.

[29] T. Oskiper, S. Samarasekera, and R. Kumar. CamSLAM: Vision aided inertial tracking and mapping framework for large scale ar applications. In *ISMAR Adjunct*, 2017.

[30] M. Palieri, B. Morrell, A. Thakur, K. Ebadi, J. Nash, A. Chatterjee, C. Kanellakis, L. Carlone, C. Guaragnella, and A. Agha-mohammad. LOCUS: A multi-sensor lidar-centric solution for high-precision odometry and 3D mapping in real-time. In *IEEE Robotics and Autonomation Letters*. IEEE, 2020.

[31] A. Rosinoi, M. Abate, Y. Chang, and L. Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[32] M. Runz and I. Agapito. Cofusion: Real-time segmentation, tracking and fusion of multiple objects. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

[33] R. Salas-Moreno, R. Newcombe, H. Strasdat, P. Kelly, and A. Davison. SLAM++: Simultaneous localisation and mapping in the level of objects. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[34] T. Shan and B. Englot. LeGO-LOAM: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018.

[35] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus. LIO-SAM: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.

[36] W. Shao, S. Vijayarangan, C. Li, and G. Kanitor. Stereo visual inertial lidar simultaneous localization and mapping. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

[37] S. Thrun, D. Hahnel, D. Ferguson, M. Montemerlo, R. Triebel, W. Burgard, C. Baker, Z. Omohundro, S. Thayer, and W. Whittaker. A system for volumetric robotic mapping of abandoned mines. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2003.

[38] J. Tian, W. Cheung, N. Glaser, Y. Liu, and Z. Kira. UNO: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[39] J. Wald, K. Tateno, J. Sturm, N. Navab, and F. Tombari. Real-time fully incremental scene understanding ob mobile platforms. *IEEE Robotics and Automation Letters*, 3(4):3402–3409, 2018.

[40] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger. MID-Fusion: Octree-based object-level multi-instance dynamic SLAM. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[41] H. Ye, Y. Chen, and M. Liu. Tightly coupled 3D lidar inertial odometry and mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[42] J. Zhang and S. Singh. LOAM: lidar odometry and mapping in real-time. *Robotics: Science and Systems*, 2(9), 2014.

[43] J. Zhang and S. Singh. Laser-visual-inertial odometry and mapping with high robustness and low drift. 35(8):1242–1264, 2018.

[44] L. Zheng, C. Zhu, J. Zhang, H. Zhao, H. Huang, M. Niessner, and K. Xu. Active scene understanding via online semantic reconstruction. *arXiv preprint arXiv:1906.07409*, 2019.

[45] R. Zlot and M. Bosse. Efficient large-scale 3D mobile mapping and surface reconstruction of an underground mine. *Field and service robotics*, 33(8):479–493, 2014.