# SubT-MRS Dataset: Pushing SLAM Towards All-weather Environments

https://superodometry.com/datasets

Supplementary Material

## A. Overview

We will provide materials from three perspectives: description and analysis of the SubT-MRS dataset, derivation of robustness metric, and more experiment results.

## B. SubT-MRS Dataset

In this section, we will provide a more detailed description of our dataset, present key statistics to highlight the challenges inherent in datasets and exhibit a gallery of ground truth 3D models from a range of diverse environments.

### B.1. Dataset Description

We presented comprehensive details about our datasets, including specific challenges, types of motion, and the sensors used, as shown in Table 1 and 2. The average ATE (Absolute Trajectory Error) is computed based on results of the top 5 teams and blue-shaded area reflects ranking of ATE.

**Overview of LiDAR Track in Real-World** Table 1 provide comprehensive dataset information from the LiDAR track, encompassing geometric degradation in real-world and simulated environments, as well as mixed degradation involving both geometric and visual factors. It's important to note that place recognition in cave settings presents challenges for backend optimization, owing to the similarity and repetitiveness of geometric features. Consequently, this results in an average ATE of 2.432 m, which is the highest.

**Overview of LiDAR Track in Simulation** The dataset was collected using a drone. Its primary challenges arose from the drone's aggressive motion with a maximum angular velocity of 360 degrees per second. Additionally, the Factory sequence, shown in Figure 6 A, was captured in snowy conditions. Such a dynamic environment can impair the performance of SLAM systems by introducing snow noise.

**Overview of LiDAR Track in Mixed Degradation** It includes both LiDAR and visual degradation such as Long Corridor and Multi-floor sequence, which will be illustrated in Sec B.2. The Block LiDAR sequence is designed to simulate sensor dropout scenarios, which are frequently encountered in the field of robotics. To evaluate the failure-aware capabilities of the systems, we intentionally disrupted the LiDAR measurements midway through the run. Our objective was to observe whether the algorithm could detect off-normal scenarios and switch to alternate modalities.

**Overview of Visual Track** Tables 2 provide comprehensive information from the visual track, including degradation from real-world and simulated environments. In real-world settings, the Low Light 1 and Low Light 2 sequences, captured in cave environments, presented place recognition challenges due to the similarity and repetitiveness of features, as shown in Figure 5 A and B. These environments look alike despite being in different locations. Figure 5 F, G, and D depict lighting changes from the Flash Light and Over Exposure sequence, which breaks the photometric consistency assumption on feature tracking. Figure 5 E depicts scenarios from the Smoke Room sequence, where the dynamic smoke not only reduces the number of detected features but also significantly increases noise in feature matching. Figure 5 H exhibits image noise in the Outdoor Night sequence, and I shows the fisheye camera covered to simulate sensor dropout scenarios. In the simulated environments, the scenes captured are shown in Figure 6 D, E, and F, respectively.

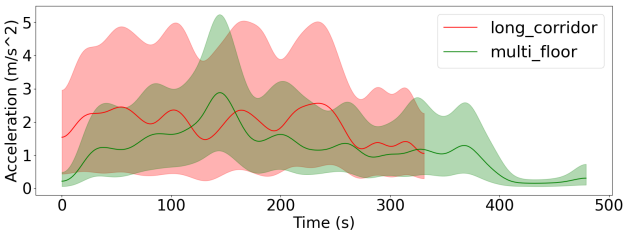


Figure 1. The graphs display average acceleration over time for all runs conducted in the long corridor and multi-floor environments. The shaded area indicates the variance in acceleration.

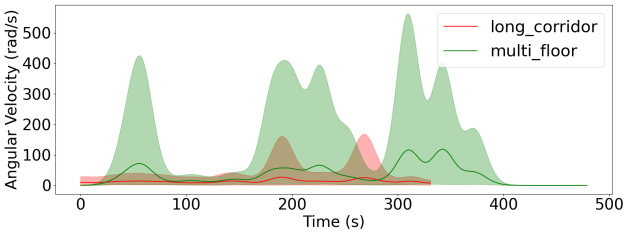


Figure 2. The graphs display average angular velocity over time for all runs conducted in the long corridor and multi-floor environments. The shaded area indicates the variance in angular velocity.

### B.2. Dataset Statistics

In this section, we will analyze our dataset from three perspectives: degeneracies in LiDAR track, degeneracies in visual track, and an analysis of robot locomotion.

**Degeneracies in LiDAR** To help users understand the challenges of our dataset, we use various metrics to measure the difficulty of dataset statistically. The statistics include

Table 1. Detailed Dataset Information on LiDAR Track (Blue shadings indicate ATE rankings)

| | | Degradation Type | | | | | | | Motion Type | | | Sensor Used | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset Seq** | | Low Light | Textureless | Structureless | Stairs | Smoke/Snow | Aggressive Motion | Repetitive Features | Vehicle Type | Max Speed | Length (m) | Fisheye | LiDAR | Thermal | IMU | **Average ATE** |
| Urban | Real World | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | UGV1 | 2 m/s | 441.86 | ✗ | ✓ | ✗ | ✓ | 0.633 |
| Tunnel | | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | UGV2 | 2 m/s | 493.67 | ✗ | ✓ | ✗ | ✓ | 0.240 |
| Cave | | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | UGV3 | 2 m/s | 593.79 | ✗ | ✓ | ✗ | ✓ | 0.696 |
| Nuclear_1 | | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | UGV1 | 2 m/s | 124.92 | ✗ | ✓ | ✗ | ✓ | 0.402 |
| Nuclear_2 | | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | UGV2 | 2m/s | 1377.37 | ✗ | ✓ | ✗ | ✓ | 0.613 |
| Laurel_Cavern | | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | Handheld | 2 m/s | 490.46 | ✓ | ✓ | ✓ | ✓ | 2.432 |
| Factory | Sim | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | Drone | 360 degree/s | 160.7 | ✗ | ✓ | ✗ | ✓ | 3.665 |
| Ocean | | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | Drone | 360 degree/s | 127.5 | ✗ | ✓ | ✗ | ✓ | 5.002 |
| Sewerage | | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | Drone | 360 degree/s | 131.0 | ✗ | ✓ | ✗ | ✓ | 7.060 |
| Long Corridor | Mix | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | RC Car | 4 m/s | 616.45 | ✓ | ✓ | ✗ | ✓ | 1.950 |
| Multi Floor | | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | Legged Robot | 2 m/s | 270 | ✓ | ✓ | ✗ | ✓ | 1.782 |
| Block LiDAR | | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | Legged Robot | 2 m/s | 307.55 | ✓ | ✓ | ✗ | ✓ | 1.099 |

Table 2. Detailed Dataset Information on Visual Track (Blue shadings indicate ATE rankings)

| | | Degradation Type | | | | | | | Motion Type | | | Sensor Used | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset Seq** | | Low Lighting | Textureless | Over Exposure | Darkness | Smoke | Aggressive Motion | Repetitive Features | Vehicle Type | Max Speed | Length (m) | Fisheye | LiDAR | Thermal | IMU | **Average ATE** |
| Low Light1 | Real World | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | Handheld | 2 m/s | 400.61 | ✓ | ✗ | ✓ | ✓ | 2.232 |
| Low Light2 | | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | Handheld | 2 m/s | 583.19 | ✓ | ✗ | ✓ | ✓ | 0.633 |
| Over Exposure | | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | Legged Robot | 2 m/s | 456.26 | ✓ | ✗ | ✓ | ✓ | 3.163 |
| Flash Light | | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | Legged Robot | 2 m/s | 147.75 | ✓ | ✗ | ✓ | ✓ | 1.476 |
| Smoke Room | | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | RC car | 2m/s | 104.84 | ✓ | ✗ | ✓ | ✓ | 4.953 |
| Outdoor Night | | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | Legged Robot | 2 m/s | 254.03 | ✓ | ✗ | ✓ | ✓ | 7.776 |
| End of World | Sim | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | Drone | 350 degree/s | 280 | ✓ | ✗ | ✗ | ✓ | 0.982 |
| Moon | | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | Drone | 60 degree/s | 850 | ✓ | ✗ | ✗ | ✓ | 8.024 |
| Western Desert | | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | Drone | 80 degree/s | 600 | ✓ | ✗ | ✗ | ✓ | 1.763 |

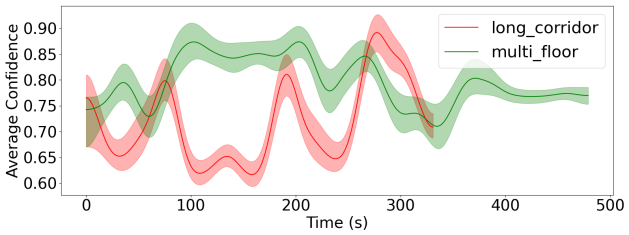


Figure 3. The graphs display the average confidence value of state estimation over time for all runs conducted in the long corridor and multi-floor environments. The shaded area indicates the variance in confidence value.

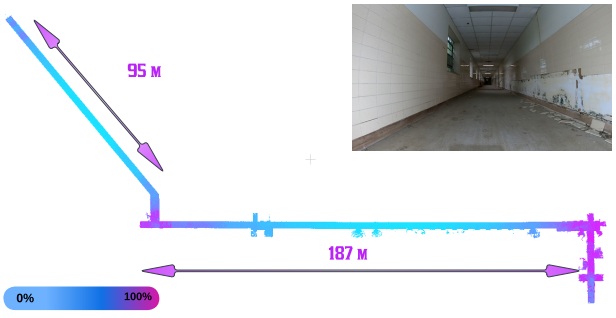


Figure 4. The Confidence Map in Long Corridor Environments. The central region of the long corridor displays the most blue areas, indicating higher uncertainty. This is primarily due to the limited constraints in the forward direction in the corridor, leading to increased mapping uncertainty.

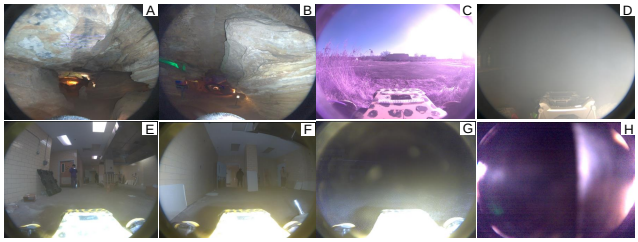


Figure 5. Real-world datasets captured with our fisheye camera to exhibit various visual degradations.

acceleration, angular velocity rate, and confidence value. The acceleration and angular velocity are used to measure the motion pattern of this dataset and a confidence value is provided by Super Odometry[**?** ] to evaluate the degradation level of environments. In the LiDAR track, we select 2 typical challenging environments: Long corridor and Multi Floor sequence to evaluate.

Figure 9 and Figure 8 demonstrate that the Long Corridor sequences exhibit greater acceleration compared to the multi-floor sequences, while the Multi-Floor sequences experience more angular velocity changes than those in long corridor environments. Given that Super Odometry can assess the confidence of state estimation, we utilize this metric to gauge the overall challenges of the dataset. It is observed that the long corridor shows lower confidence values most of the time, indicating that it is a more challenging environment for LiDAR SLAM systems. Figure 4 further illustrated the confidence value of map in 3D space.

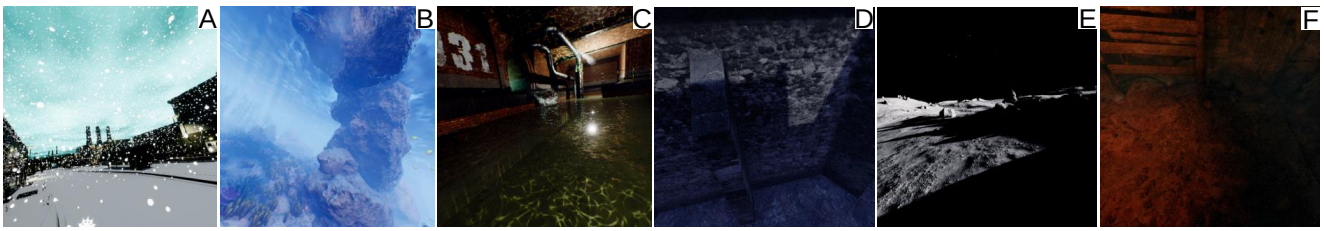


Figure 6. Simulation from both LiDAR/Visual Track. A: Factory B: Ocean C: Swerage D: End of world E: Moon F: Western Desert

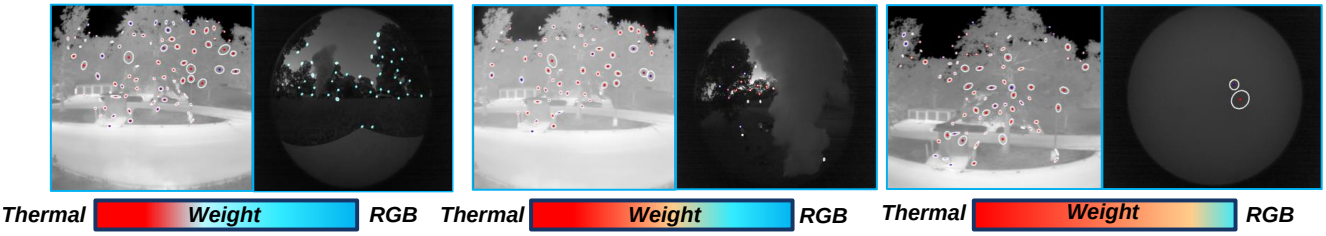


Figure 7. Feature Tracking in Smoke Environments on RGB and Thermal Images: The sequence from left to right on weight illustrates a decrease in feature tracking on RGB images and a corresponding increase on thermal images.

**Degeneracies in Visual** We selected a smoke environment to specifically illustrate degeneracy, showcasing the tracked feature count in both thermal and RGB images, as shown in Figure 7. In this setting, we applied a learning-based feature extraction pipeline[**?** ], to these multispectral images. With increasing smoke density, the thermal images are not influenced, whereas RGB images tend to add noise to the feature tracking process. Therefore, in such scenarios, an optimal SLAM solution should gradually prioritize the thermal camera as primary sensor for state estimation.

**Heterogeneous Robot Locomotion** Our dataset, sourced from heterogeneous robots, allows for an analysis of their movement patterns to identify which robots yield the more challenging datasets. Figure 9 and Figure 8 present the average acceleration and angular velocity across all runs from different robots. These figures highlight that the dataset from the canary drone poses greater challenges compared to others, evidenced by its higher frequency of peaks in both acceleration and angular velocity.

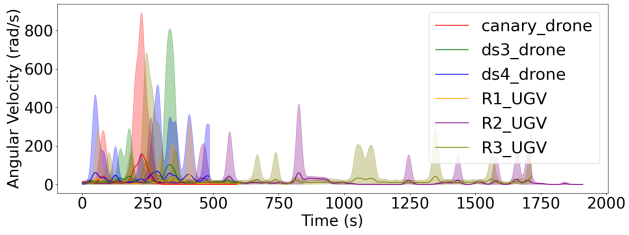


Figure 8. The graph displays the angular velocity over time for all runs conducted in the SubT environment. The shaded area indicates the variance in angular velocity.

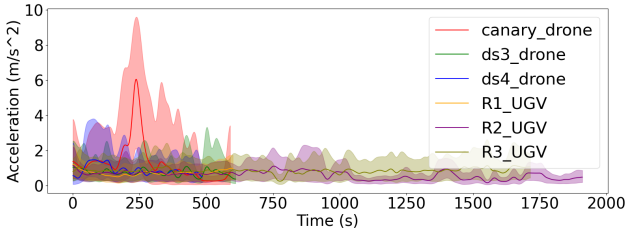


Figure 9. The graph displays the acceleration over time for all runs conducted in the SubT environment. The shaded area indicates the variance in acceleration.

### B.3. Ground Truth Model

In this section, we are pleased to present our extensive collection of Ground Truth 3D Models, as displayed in Figure 10. This collection includes a ground truth model of a university campus, covering both indoor and outdoor areas, extensive loop mapping for long corridor and multi floor environments. Importantly, all ground truth models were captured using a FARO scanner, guaranteeing an accuracy within 10 cm.

## C. Derivation of Robustness Metric

As introduced in the main paper Sec 3.3, we describe how to use B-spline to derive estimated linear and angular velocity. Additionally, we will show the steps of robustness metric calculation.

**B-splines** We adopt B-splines [**?** ] to interpolate the given trajectory, yielding continuous-time trajectories in $SE(3)$. The continuous trajectory is decided by control poses $T_{w,i}$, where $T_{w,i}$ is the estimated poses at time $t_i$ in a world co-

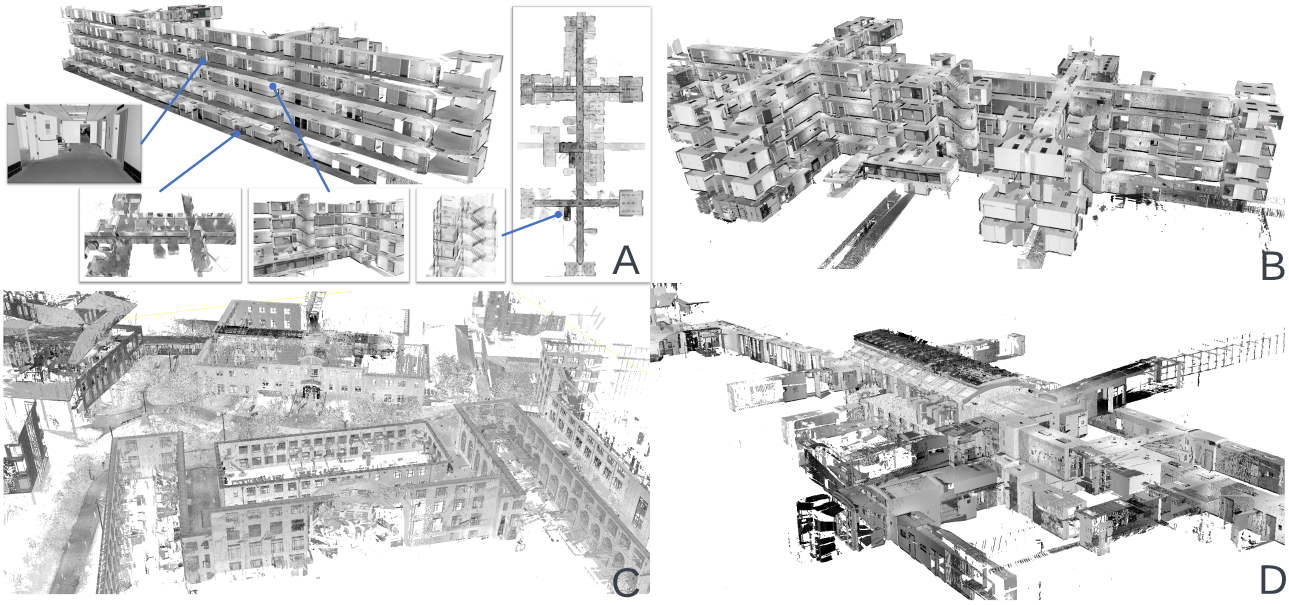


Figure 10. The Ground Truth Model in SubT-MRS Dataset. A: Long Corridor B: MultiFloor C: Campus Outdoor D: Campus Indoor

ordinate system $w$. According to the locality of the cubic B-spline basis, the value of the spline curve at any time $t$ is decided by four control poses. We use one absolute pose, $T_{w,i-1}$, and three incremental poses, parametrized by twists $\xi_i$. More specifically, the spline trajectory is given by

$$T_{w,s}(u(t)) = T_{w,i-1} \prod_{j=1}^{3} \exp(\mathbf{B}_j(u(t))\xi_{i+j-1}) \quad (1)$$

where the exp denotes the matrix exponential. $u(t) = (t - t_i)/\Delta t \in [0, 1]$ and $t_i = i\Delta t$ are used in the cumulative basis functions for the B-splines.

$$\bar{\mathbf{B}}(u) = C \begin{bmatrix} 1 \\ u \\ u^2 \\ u^3 \end{bmatrix}, \quad C = \begin{bmatrix} 6 & 0 & 0 & 0 \\ 5 & 3 & -3 & 1 \\ 1 & 3 & 3 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (2)$$

In Eq.1, $\mathbf{B}_j$ is the $j$-th entry (0-based) of vector $\mathbf{B}$. The incremental pose from time $t_{i-1}$ to $t_i$ is encoded by twist

$$\xi_i = \log(T_{w,i-1}^{-1} T_{w,i}), \quad (3)$$

$$\frac{\partial}{\partial t_s} T_{w,s}(u(t)) = [\Delta\dot{R}, \Delta\dot{T}] = [\omega_w, \mathbf{v}_w] \quad (4)$$

We differentiate the spline trajectory as presented in Eq. 4 to obtain estimates of the velocity, denoted as $\mathbf{v}_w$, and the angular velocity, denoted as $\omega_w$. We sampled velocities from both the interpolated ground truth trajectory and the trajectory estimated via B-spline, establishing velocity correspondences based on time stamps.

**Robustness Metric Calculation** We introduce the F-score [**?** ] to evaluate the robustness of SLAM algorithms. The robustness metric can be calculated in the following steps:

**Algorithm 1** Robustness Metric Calculation

**Input:** Trajectory $T_{w,i}$, ground truth velocities $\mathbf{v}_{gt}^w, \boldsymbol{\omega}_{gt}^w$
**Output:** Robustness Metric
**Step 1:** Smooth trajectory with B-spline on $T_{w,i}$
**Step 2:** Derive $\mathbf{v}_{est}^w, \omega_{est}^w$ from smoothed trajectory
**Step 3:** Calculate distances: $\mathbf{v}_{dis} = ||\mathbf{v}_{gt}^w - \mathbf{v}_{est}^w||$, $\boldsymbol{\omega}_{dis} = ||\boldsymbol{\omega}_{gt}^w - \boldsymbol{\omega}_{est}^w||$
**Step 4:** Compute F-scores $F(\mathbf{v}_{dis}), F(\boldsymbol{\omega}_{dis})$ for various thresholds
**Step 5:** Plot F-scores against thresholds; area under curve is robustness metric

## D. Experiments

In this section, we provide detailed experiments assessing accuracy and robustness, which were not included in the main paper due to length constraints. It should be noted that the main paper already includes comprehensive experiments on our datasets and presents summarized conclusions. The additional experiments here serve as **supplementary evidence supporting the conclusions** drawn in Sections 4.1 and 4.2 of our paper.

### D.1. Accuracy Evaluation

In section 4.1, we presented Absolute Trajectory Error (ATE) values for the LiDAR and visual tracks. We omit-

Table 3. Accuracy Performance on LiDAR Degradation. Red numbers represent ATE/RPE ranking. * denotes incomplete submissions.

| | | Real World | | | | | | Simulation | | | Mix Degradation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Team** | | Urban | Tunnel | Cave | Nuclear_1 | Nuclear_2 | Laurel_Caverns | Factory | Ocean | Sewerage | Long Corridor | Multi Floor | Block Lidar | **Average** |
| Liu et al[1] | ATE | 0.307 | 0.095 | 0.629 | 0.122 | 0.235 | **0.260** | **0.889** | 0.757 | **0.978** | 1.454 | **0.401** | **0.934** | **0.588** |
| Yibin et al[4] | ATE | 1.060 | 0.220 | 0.750 | 0.470 | 0.620 | 9.140 | 4.920 | **0.280** | 24.460 | 2.990 | 5.500 | 1.340 | 4.312 |
| Weitong et al[2] | ATE | **0.26** | 0.096 | **0.617** | **0.120** | **0.222** | 0.402 | 0.998 | 0.770 | 1.586 | **1.254** | 0.577 | 1.056 | 0.663 |
| Kim et al[3] | ATE | 0.331 | **0.092** | 0.787 | 0.123 | 0.270 | 0.279 | 10.628 | 22.425 | 7.147 | 2.100 | 0.650 | 1.068 | 3.825 |
| Zhong et al[5] | ATE | 1.205 | 0.695 | - | 1.175 | 1.72 | 2.08 | **0.889** | 0.778 | 1.13 | - | - | - | 1.209* |
| Liu et al[1] | RPE | 0.038 | **0.032** | **0.055** | **0.028** | **0.048** | **0.040** | 0.191 | 0.174 | 0.188 | 0.088 | 0.059 | **0.148** | **0.091** |
| Yibin et al[3] | RPE | 0.130 | 0.090 | 0.150 | 0.130 | 0.200 | 0.200 | **0.040** | **0.040** | **0.160** | 0.630 | 0.280 | 0.180 | 0.186 |
| Weitong et al[2] | RPE | **0.038** | **0.032** | 0.056 | 0.029 | 0.049 | 0.046 | 0.190 | 0.183 | 0.243 | **0.086** | **0.054** | 0.166 | 0.097 |
| Kim et al[4] | RPE | 0.098 | **0.032** | 0.055 | **0.028** | 0.054 | 0.046 | 0.861 | 0.535 | 0.401 | 0.093 | 0.26 | 0.167 | 0.219 |
| Zhong et al[5] | RPE | 0.157 | 0.062 | - | 0.079 | 0.1062 | 0.0937 | 0.706 | 0.691 | 0.617 | - | - | - | 0.313* |

Table 4. Accuracy Performance on Visual Degradation. Red numbers represent ATE/RPE ranking. * denotes incomplete submissions.

| | | Real World | | | | | | Simulation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Team** | | Lowlight 1 | Lowlight 2 | Over Exposure | Flash Light | Smoke Room | Outdoor Night | End of World | Moon | Western Desert | **Average** |
| Peng et al[1] | ATE | 1.063 | 1.637 | **0.503** | **0.44** | **0.153** | **0.827** | **0.038** | **0.195** | **0.070** | **0.547** |
| Jiang et al[3] | ATE | **1.019** | **1.126** | 1.911 | 2.341 | 3.757 | 11.821 | 2.154 | 0.604 | 4.010 | 3.193 |
| Thien et al[2] | ATE | 1.081 | 2.054 | 1.733 | 1.054 | 10.532 | 7.692 | 0.753 | 1.228 | 1.209 | 3.037 |
| Li et al[4] | ATE | 5.768 | 7.834 | 1.757 | 1.295 | 5.370 | 10.766 | - | 30.07 | - | 8.98* |
| Peng et al[1] | RPE | **0.058** | **0.063** | **0.051** | **0.149** | **0.026** | **0.064** | **0.002** | **0.014** | **0.01** | **0.048** |
| Jiang et al[3] | RPE | 0.190 | 0.203 | 0.258 | 0.307 | 0.884 | 3.427 | 3.982 | 0.792 | 7.477 | 1.947 |
| Thien et al[2] | RPE | 0.197 | 0.186 | 0.181 | 0.231 | 0.071 | 0.279 | 0.471 | 0.007 | 0.777 | 0.266 |
| Li et al[4] | RPE | 0.088 | 0.088 | 0.124 | 0.160 | 0.911 | 0.478 | - | 0.347 | - | 0.314* |

ted the RPE results due to the constraints of paper length. Table 3 and Table 4 provide RPE results.

### D.2. Robustness Evaluation

In section 4.2, we presented a summary of the robustness metrics for the LiDAR and visual tracks and omitted the detailed robustness plots for each sequence due to the constraints of the paper's length. Figure 11 and Figure 12 illustrate the robustness metrics on each sequence for the LiDAR teams, while Figure 13 show the results from visual track teams.

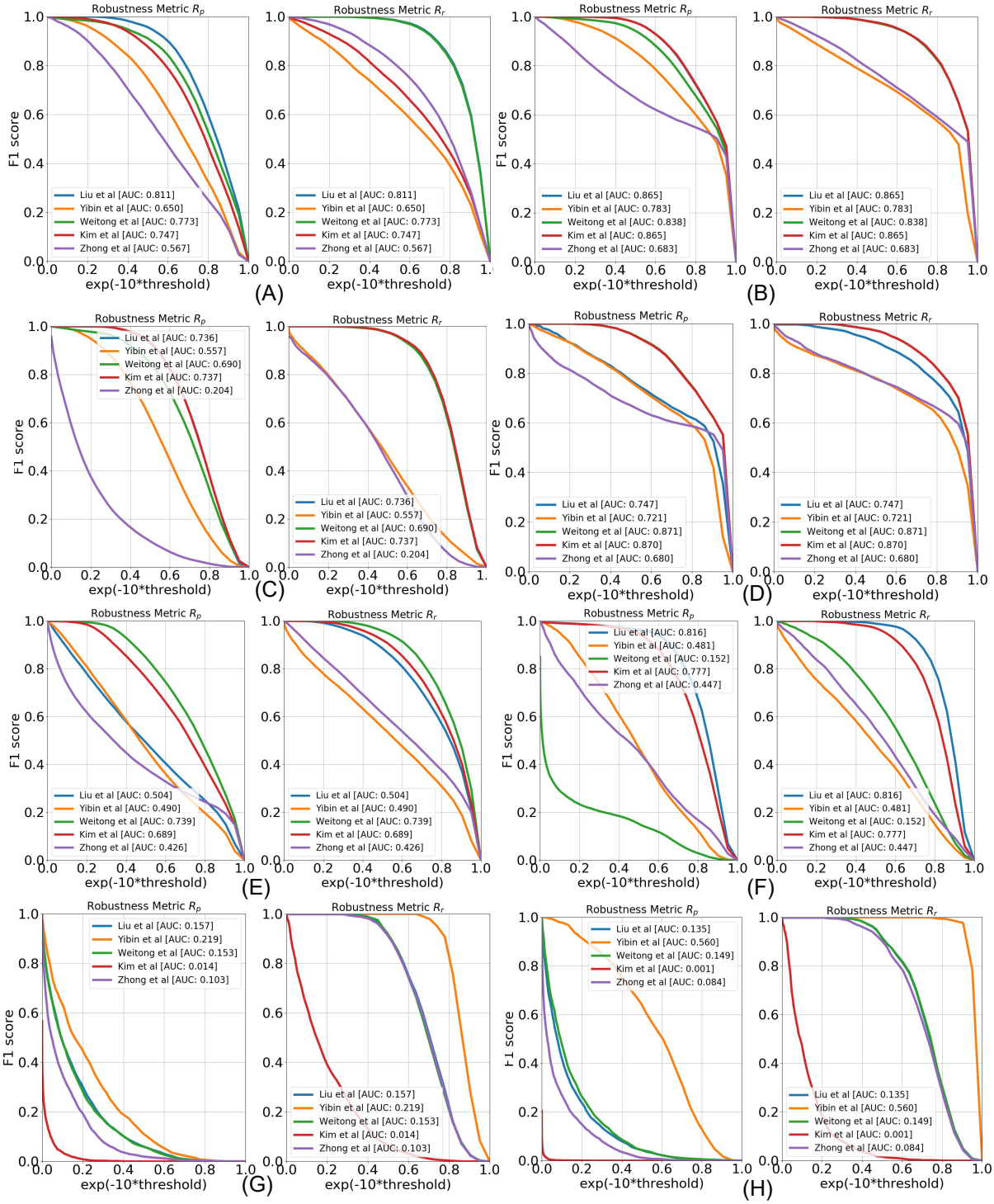


Figure 11. The image, arranged from left to right, displays the robustness metrics $R_p$ and $R_r$ from LiDAR track teams: (A) Urban, (B) Tunnel, (C) Cave, and (D) Nuclear_1 (E) Nuclear_2 (F) Laurel Cavern (G) Factory (H) Ocean

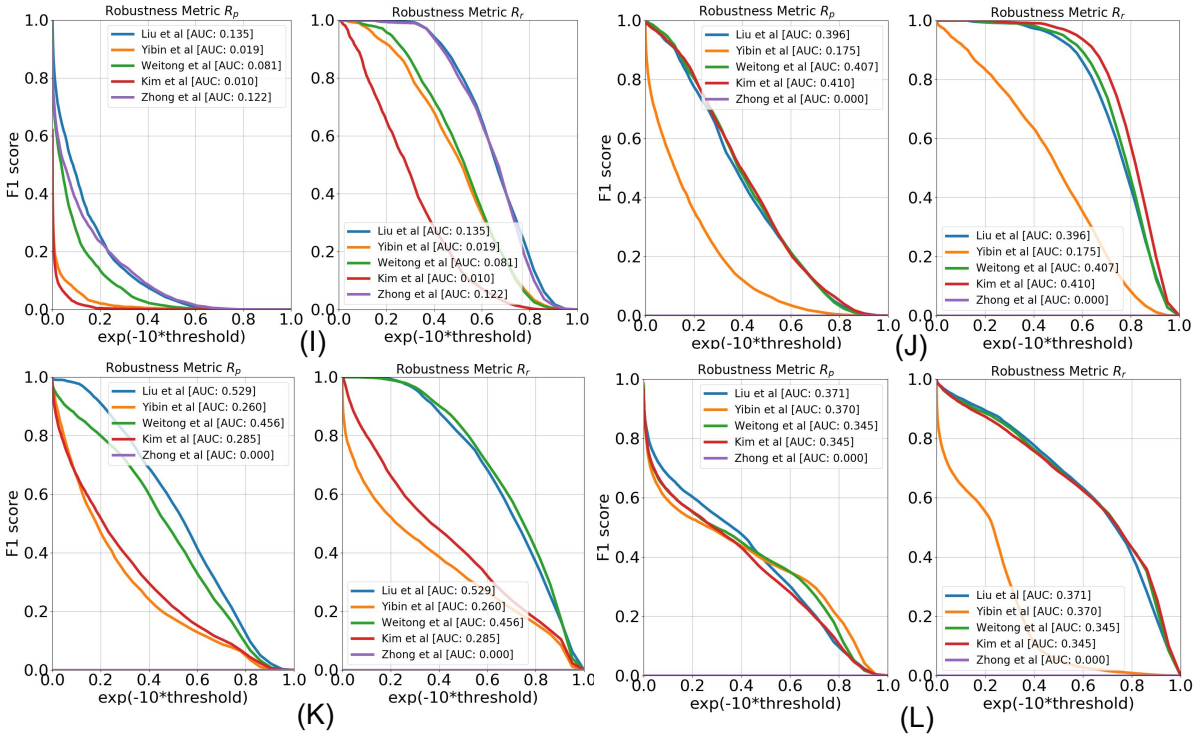


Figure 12. The image, arranged from left to right, displays the robustness metrics $R_p$ and $R_r$ from LiDAR track teams: (I) Sewerage, (J) Long Corridor, (K) Multi Floor, (L) Block LiDAR

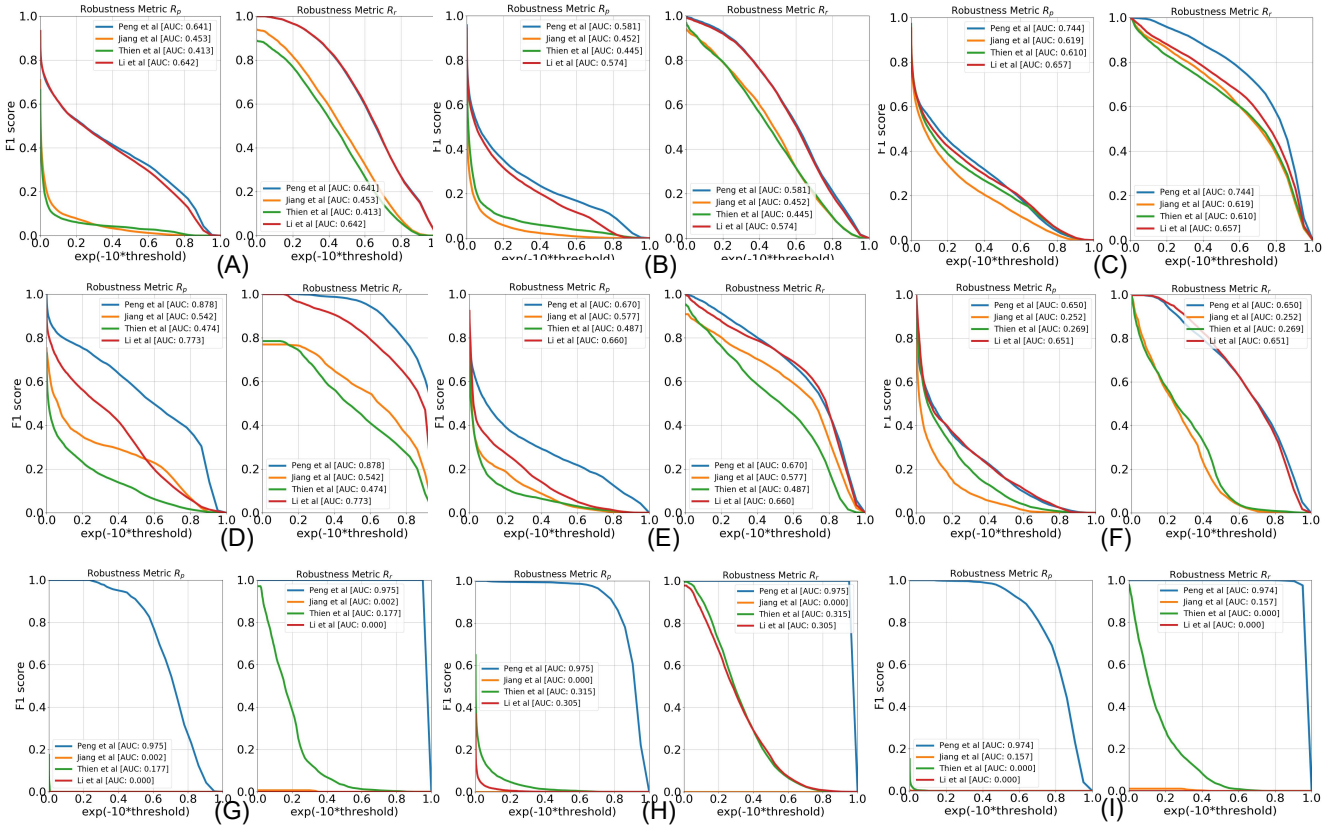


Figure 13. The image, arranged from left to right, displays the robustness metrics $R_p$ and $R_r$ from visual track teams: (A) Low Light1, (B) Low Light2, (C) Over Exposure, and (D) Flash Light,(E) Smoke Room (F) Outdoor Night, (G) End of World, (H) Moon (I) Western Desert