2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Robot Traversability Prediction: Towards Third-Person-View Extension of Walk2Map with Photometric and Physical Constraints

Jonathan Tay Yu Liang Kanji Tanaka

***Abstract*— Walk2Map has emerged as a promising data-driven method to generate indoor traversability maps based solely on pedestrian trajectories, offering great potential for indoor robot navigation. In this study, we investigate a novel approach called Walk2Map++, which involves replacing Walk2Map's first-person sensor (i.e., IMU) with a human observing third-person view from the robot's onboard camera. However, human observation from a third-person camera is significantly ill-posed due to visual uncertainties resulting from occlusion, nonlinear perspective, depth ambiguity, and human-to-human interaction. To regularize the ill-posedness, we propose integrating two types of constraints: photometric (i.e., occlusion ordering) and physical (i.e., collision avoidance). We demonstrate that these constraints can be effectively inferred from the interaction between past and present observations, human trackers, and object reconstructions. We depict the seamless integration of asynchronous map optimization events, like loop closure, into the real-time traversability map, facilitating incremental and efficient map refinement. We validate the efficacy of our enhanced methodology through rigorous fusion and comparison with established techniques, demonstrating its capability to advance traversability prediction in complex indoor environments. The code and datasets associated with this study are available for further research and adoption in the field at `https://github.com/jonathanty197/HO3-SLAM`.**

I. Introduction

Traversability prediction, which involves visually recognizing whether a particular area on a 2D floor is navigable or not, presents a fundamental challenge in visual robot navigation. This problem has been approached through various formulations, such as region-wise traversability prediction [1]–[13]. However, most existing studies have focused on outdoors with good visibility, such as grasslands or corridors. In contrast, in crowded, occlusion-prone environments such as offices, these techniques are severely limited in their applicability. Figure 1 shows an example of such a crowded office, where the floor area is difficult to see directly.

Several recent studies have reported traversability studies targeting office environments. Kucner et al. [14] focused on constructing maps of dynamics (MoDs), which encode semantic information regarding typical motion patterns within a given environment using laser range finders. Meanwhile, Alempijevic et al. [15] addressed the mapping of human motion dynamics by utilizing interactions with humans to identify areas where traversability changes occur. Additionally, Papadakis et al. [16] proposed a generative methodology

Fig. 1. Vision-based traversability prediction is an open problem in crowded office environments where occlusions and obstacles are rich. In a crowded dynamic scene, it is difficult to obtain a good point cloud map and human detection due to occlusions and obstacles. The left and right panels show the DSO point cloud map and Detectron2 human detection mask, respectively.

to enhance mobile robot mapping and navigation in indoor environments, leveraging human spatial activity for passage detection and occupancy prediction, while effectively mitigating false positive human detections using prior map information. Nevertheless, its focus is primarily on scenarios characterized by minimal obstacles and occlusions, leaving a significant gap in addressing challenges prevalent in typical office environments. The broader issue of navigating through general office spaces remains largely unresolved.

To address the issue of obstacles and occlusions in a scene, a new data-driven approach to constructing floor plans known as Walk2Map [17] has recently emerged and gained traction. This method offers a simple yet powerful means of generating floor plans based solely on the trajectories of individuals walking indoors. The advancements inspire it in affordable and high-performance equipment, such as IMU measurements on smartphones and data-driven inertial odometry. Remarkably, the floor plans generated through Walk2Map exhibit exceptional quality, it can be seen that it has ideal qualities as a traversability map [18] for mobile robots. However, the use of this floor plan as a traversability map for autonomous mobile robot applications is severely limited by the requirement to equip pedestrians with odometers. Overcoming this hurdle is a key focus of our research.

In this study, we investigate a novel approach called Walk2Map++, which involves replacing Walk2Map's first-person sensor (i.e., IMU) with a human observing third-person view from the robot's onboard camera. This innovative approach offers several advantages. Firstly, it enables

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 20K12008 and 23K11270.

∗J. T. Y. Liang and K. Tanaka are with Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering, University of Fukui, Japan. {mf228029@g., tnkknj@}u-fukui.ac.jp

the acquisition of the traversability map in a robot-centric coordinate system as opposed to a pedestrian-centric one, facilitating long-horizon path-planning by robots. Moreover, measurement data from multiple pedestrians can be collected in real-time without relying on wireless networks, hence the efficiency of map construction is enhanced. In this case, our approach shares the same objective with dynamic map-building approaches such as DynaSLAM [19] in that it uses people observation and obstacle reconstruction to build dynamic environment maps. Still, the key difference is that we focus on the efficient building of a traversability map rather than 3D point cloud maps.

To address the problem's ill-posed nature, we leverage photometric and physical constraints, focusing on human-object occlusion ordering and collision avoidance. These constraints are inferred from observations using object reconstruction and human trackers. We also demonstrate real-time integration of asynchronous map optimization for incremental traversability map refinement. Our methodology is compared with established techniques, validating its effectiveness in improving traversability prediction in complex environments.

The contributions of this paper are as follows: (i) We address Walk2Map++, a challenging problem that replaces Walk2Map's first-person sensor (i.e. IMU) with a human observing third-person view from a robot's onboard camera. (ii) Establishing dynamic relationships and physical constraints between moving humans and stationary objects through the utilization of SLAM and human detectors, which enables the observation of interactions within the environment. (iii) Employing photometric constraints to determine dynamic human locations in the transition from 3D to 2D space through a novel occlusion ordering algorithm and human pose estimation methods, and further enhancing the understanding of human positions in complex scenarios. (iv) Evaluation of the method's effectiveness through a concrete performance index for traversability maps, validated via fusion and comparison with well-known methods in comprehensive real-world experiments. (v) The code and datasets associated with this study are publicly available[1], facilitating further research and adoption in the field.

II. Related Works

A. Traversability Prediction

Traversability prediction, a prominent facet of computer vision research, has undergone notable advancements within the supervised learning paradigm [20]. Recent methodologies exhibit the evolving landscape of this field, expanding its scope from outdoor terrains to intricate indoor environments. Researchers have explored the effectiveness of generating control commands based on onboard sensor data, demonstrating efficacy in predicting traversable regions within complex indoor spaces [21]. A noteworthy trend involves the exploration of self-supervised frameworks for autonomous robot applications, addressing challenges such as long-range traversability [22], RGB-D traversability prediction [23], visibility challenging environments [24], and hazardous forest scenarios [25]. These frameworks capitalize on the synergistic use of Simultaneous Localization and Mapping (SLAM) and Multi-Object Tracking (MOT), monitoring both stationary objects and moving entities in real-time to enhance overall traversability prediction. In addition to supervised learning, emerging semi-supervised [26] and unsupervised [27] frameworks, leveraging scene geometry, appearance, and range-color information, show promise. However, these existing methods do not assume crowded or occlusion-rich environments and do not provide effective clues to the floor area in these challenging environments. In contrast, our approach allows the emphasis on predicting traversable areas within occluded regions, a critical aspect for navigating challenging terrains with restricted visibility, marking a significant stride toward the development of autonomous robot systems adept at handling complex environments.

B. Human Moving Trails Observation

In addressing the enduring challenge of scene arrangement recovery under moderate to heavy occlusion in monocular video analysis, Monszpart et. al. introduce iMapper [28], a data-driven method that uniquely leverages the correlation between human-object interactions and scene-object arrangements. By identifying characteristic interactions and employing an innovative occlusion-aware matching procedure, iMapper yields substantial advancements in both scene analysis and 3D human pose recovery, particularly in scenarios with medium to heavy occlusion, as demonstrated through rigorous quantitative and qualitative evaluations. The idea of creating maps from human observations in a fixed camera or non-occluded setup is not new [29]. Our main difference is that we use a moving camera and assume a crowded environment with rich occlusions and obstacles.

On the other hand, Walk2Map [17] is a data-driven approach for constructing floor plans solely from the trajectories of people walking indoors. It leverages the movements of individuals equipped with ego-motion sensors, such as IMU (Inertial Measurement Unit) measurements on smartphones, to generate high-quality floor plans. We observe that these floor plans are of good quality and could be used as traversability maps for indoor robot navigation. However, Walk2Map assumes that a first-person sensor such as an IMU will be attached to a human, and is not intended for use in autonomous mobile robots. In contrast, in this study, we wish to achieve the same functionality (Walk2Map++) by not relying on that premise and using the robot's third-person camera as the sole sensing device.

III. Approach

We present an overview of our approach in Figure 2, which provides an overview of the proposed Walk2Map++ approach and the whole framework. Walk2Map++ takes image sequences as input and outputs a traversability map (Section III-A). Walk2Map++ aims to realize the functionality of

[1]https://github.com/jonathanty197/HO3-SLAM

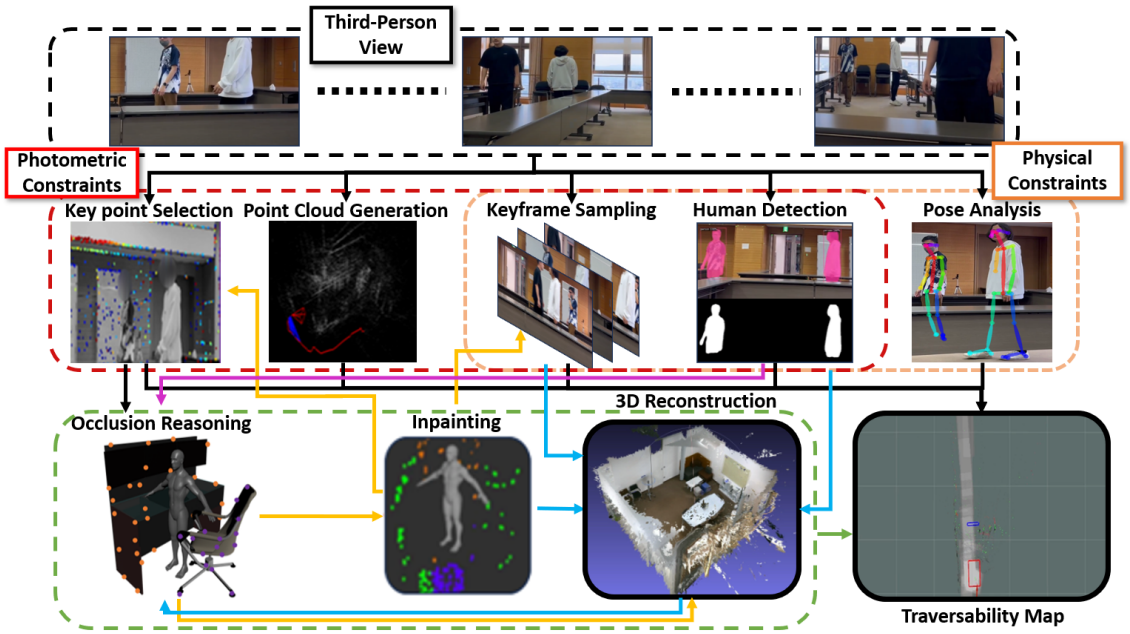


Fig. 2. Block diagram of framework: All modules are interconnected via ROS (Robot Operating System). The PHYS module comprises DSO (Direct Sparse Odometry) and is responsible for generating point clouds utilized by other modules. Within the PHOT module, a Human-Object Occlusion Ordering Algorithm is employed to extract occlusion ordering information, which is then combined with point cloud coordinates derived from Detectron2 human masks. Additionally, the Walk2Map++ module utilizes human pose estimation to predict human distance from the camera and estimate traversable regions. These traversability maps are visualized using the rviz visualizer. In the traversability map image, the red box represent the estimated human location, hence the traversable region. The grey path indicates the traversable region, which has been walked by the human. The grey path is inprinted by the red boxes.

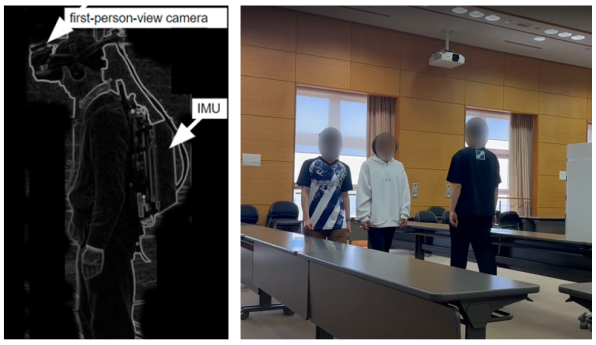


Fig. 3. Traversability prediction under severe occlusion. Left: Conventional first-person-view setup with IMU. Right: Proposed third-person-view monocular vision setup.

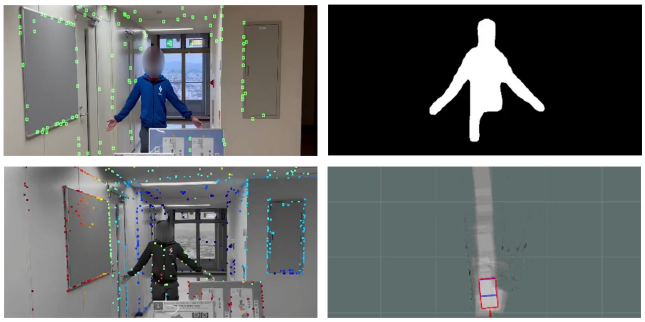


Fig. 4. Top-left: The projection of point clouds to the keyframes visualize; Top-right: Human mask used for occlusion ordering algorithm; Bottom-left: DSO; Bottom-right: Traversability map visualized with rviz visualizer.

Walk2Map using a third-person robot view instead of a first-person view human equipped with an IMU sensor, generating a traversability map in grid map format solely from the trajectories of people walking indoors, as shown in Figure 3. This module restricts human locations by analyzing human behavior and makes the traversability map as accurate as possible (Section III-B). However, this base method alone may not provide sufficient performance under crowded and occlusion situations. Two more modules will be introduced for augmentation. Primarily, leveraging the physical constraint of humans avoiding collisions with obstacles, we constrain human behavior towards incremental obstacle maps to enhance the accuracy of the traversability map (Section III-C). Secondly, based on the occlusion ordering between humans and obstacles, a photometric constraint is introduced to constrain the depth ordering of humans and obstacles from the camera's viewpoint to further enhance the accuracy of the traversability map (Section III-D). Additionally, incremental map updates are supported for asynchronous map optimization during loop closing (Section III-E). Figure 4 shows the different modules of our framework.

A. Traversability Map

The traversability map constitutes a vital component within the operational framework of the robotic system, manifesting as a two-dimensional grid overlaying the mobile plane of the robot. The grid, characterized by its discrete partitioning into cells, adheres to a spatial resolution of 10 cm $\times$ 10 cm, meticulously structured to facilitate precise navigation. Through extensive analysis, it has been determined that

the adoption of a finer cell granularity yields only marginal enhancements in performance, while significantly amplifying both computational overheads and storage requisites.

Each cell within the grid is endowed with the capacity to assume one of three distinct states:"traversable," denoting navigable terrain; "untraversable," indicative of impassable regions; and "unknown," signifying areas yet to be surveyed or categorized. In the initialization phase, all grid cells uniformly commence with an initial state of "unknown," awaiting subsequent evaluation and classification.

$$T_{\text{combined}} = T_1 \cap T_2 \tag{1}$$

Finally, we update the traversability map based on the obstacle map using Equation 1, where $T_{combined}$ signifies the common traversable area that is present in both T_1 and T_2. T_1 indicates the traversable map created by Walk2Map++, a human pose distance estimation method (Section III-B), while T_2 indicates the traversable map created by photometric constraints using a novel human-object occlusion ordering algorithm (Section III-D).

B. Walk2Map++

Walk2Map++ endeavors to accurately predict the position of humans relative to the camera by analyzing their heads and posture. Building upon prior research on pedestrian posture analysis, it has been observed that the length of a pedestrian's torso remains relatively constant during walking, providing valuable insight for position analysis.

$$D = k \cdot \frac{f \cdot L}{\|P_1 - P_2\|} \tag{2}$$

Specifically, utilizing the measured torso length denoted as L and the distance from the camera to the human torso denoted as D. f is the focal length of the camera and $\|P_1 - P_2\|$ represents the Euclidean distance between the two torso keypoints. The relationship is established through Equation 2, where Euclidean distance calculation is employed. In this study, we utilize the state-of-the-art human pose estimation model - OpenPose [30], for this purpose. The human pose tracking model used is a pre-trained lightweight model.

Prior to model implementation, parameter calibration is imperative, with a focus on determining the key parameter k in Equation 2. k is the calibration constant which helps to account for any discrepancies between the ideal pinhole model and the actual camera system. Utilizing two keypoints from human pose estimation, such as those illustrated in Figure 5, allows for the inference of torso length in pixel units. However, it's important to note that the 3D coordinates do not directly translate to physical distances in meters, necessitating the application of a scale factor. Additionally, accurate measurements based on known distances or sizes of objects within the scene are essential for the precise determination of D.

This torso-keypoint-based method demonstrates robustness compared to alternative approaches, as it does not require pedestrians to consistently face the camera for key points acquisition; rather, keypoints data can be obtained as long as the pedestrian is within the camera's view. Therefore, the algorithm is deemed valid at all angles of observation.

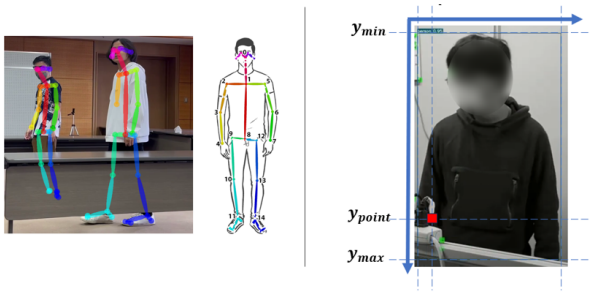


Fig. 5. (Left, Middle): Human-centric coordinate system. As shown in the middle figure, keypoint 1 and keypoint 8 is used as a reference point of torso length.(Right): The relationship between occluder's feature point and occluded human's region.

C. Physical Constraints (PHYS)

Motivated by the fact that obstacle areas are more frequently visible than traversable floor areas, we construct an online obstacle map and simultaneously constrain the position of the human using physical constraints of humans avoiding collisions with obstacles. For this purpose, we employ DSO [31] because this specific SLAM algorithm provides high discrimination ability between dynamic and static objects, yielding a reasonably dense point cloud format obstacle map. Moreover, it is common for floor and ceiling areas to be omitted from the obstacle map due to occlusion and it's necessary to note that obstacles positioned higher than human height do not pose physical constraints on humans. Therefore, specifically, through the following steps, we generate a two-dimensional high-confidence obstacle map: (1) Using the torso key points from human pose estimation, the estimated distance of humans from the camera is obtained, and then translated into 2D space of traversability map. The estimated human location area is painted and defined as "Traversable" in the traversability map. (2) DSO point clouds that are on the ceiling, floor, and too far from the camera are filtered out. (3) The traversable area from both human pose estimation and occlusion ordering algorithm is found. The intersection between both traversability maps is updated into the final map.

D. Photometric Constraints (PHOT)

Occlusion serves as both a hindrance to the recognition of third-person viewpoint cameras and a cue for determining the depth order between humans and obstacles. By combining information from the human masks obtained through Detectron2 [32] and the point cloud extracted from DSO, we have developed an occlusion ordering algorithm. We utilize a pre-trained Detectron2 model to generate accurate human masks from the camera frames, and these masks are then combined with the point cloud obtained from DSO to create a fused representation. This algorithm analyzes the fused information to determine the occlusion order between humans and objects in the scene, aiding in more accurate scene understanding.

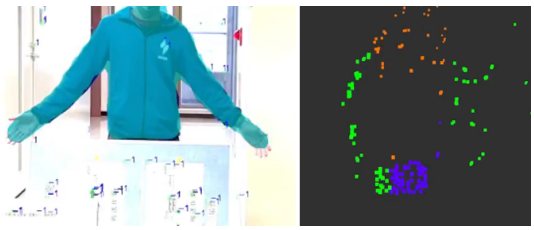


Fig. 6. Left: Occlusion ordering algorithm; Right: Grouped cluster point cloud visualized in rviz visualizer. Purple cluster indicate points that are in front of the human, orange cluster indicate points that are behind of the human.

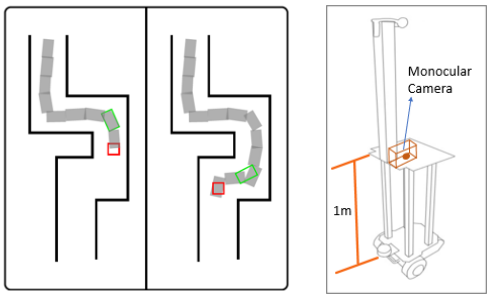
The algorithm is employed to determine whether the points are positioned in front of or behind the human with a higher likelihood.

As shown in Figure 5, the red point represents the point with coordinates (x_{point} and y_{point}), while (x_{min}, x_{max}, y_{min}, and y_{max}) represents the x and y range of the human area. Armed with this information, we can subsequently infer that pedestrians are likely situated amidst clusters of point clouds, indicating a higher probability of traversability within the intermediate area. If x_{point} and y_{point} are in x and y range, then the point is in front of the human with a higher probability. Each cell within the grid is endowed with the capacity to assume one of three distinct states: "traversable", denoting navigable terrain; "untraversable", an indicator of impassable regions; and "unknown", signifying areas yet to be surveyed or categorized. In this study, the values assigned to the points are subsequently reflected on the traversability map, with clusters of points being classified and grouped. The grid map contains values ranging from 0 to 255. Traversable areas are designated by a decrease of 5, while obstacles are marked by an increase of 10. From these clusters, we can deduce the potential location of a human situated between them, as illustrated in Figure 6.

It is essential to underscore that our occlusion ordering algorithm is grounded in a specific assumption: the upper body of the pedestrian is frequently unoccluded. While this assumption may not be extensively discussed in the current context, it is crucial to highlight that, in practical scenarios, this assumption often proves accurate and is not deemed excessively restrictive. However, it is imperative to acknowledge that the occlusion ordering algorithm becomes ineffective when the human body is fully occluded or when most of the body is occluded. In such cases, these instances will be disregarded, and the algorithm will wait until the upper body of the human becomes visible again.

E. Asynchronous Map Fusion

In order for the traversability map reconstruction task to function as an add-on to an existing online SLAM system and support incremental map construction, we implement the ability to dynamically update optimized traversability maps to properly reflect various asynchronous map optimization events, such as SLAM loop closures and map merging. Our devised framework seamlessly integrates modules, DSO [31], Human pose estimation for human distance estimation, and human occlusion ordering algorithm.



(a) (b)

Fig. 7. (a) Online Traversability Map: In our traversability map, the red box represents the current camera position while the green box represents the estimated human location. The gray region indicates the human trail, hence the traversable region. The black lines indicates the obstacles or point cloud clusters. (b) Observer Robot set up, with a right-facing monocular camera mounted on the platform of approx. 1m height from ground

Walk2Map++, PHYS and PHOT mentioned so far all provide a traversability area relative to the robot pose, so they are valid even if the robot pose is modified via map optimization events, there is no need for re-computation. To generate a sparse point cloud map, a video stream is fed into DSO. Concurrently, these points are projected onto keyframes. The occlusion ordering algorithm, in tandem with the Detectron2 [32] human mask and YOLOv7 [33] human detection model, discerns the occlusion hierarchy of humans and point cloud. Detectron2's detection outcomes, represented by bounding boxes, ascertain the location of the human body, informing both human pose estimation and occlusion ordering algorithm modules to predict the human's location and consequently determine the traversable region. The final prediction of the traversable region is the intersection of these two areas, culminating in a cohesive representation within the complete framework, as shown in Figure 7(a).

For PHYS, the traversability map can be generated on the fly by considering the point cloud from the DSO map as obstacles (Untraversable areas). In PHOT, the robot's viewpoint provides crucial information for the traversability map, with the relative position of the human observed in each frame. Additionally, the traversability map is continuously updated on the fly.

IV. Experimental Results

A. Data Preparation

We conducted an extensive data collection process to curate three distinct datasets, each comprising multiple human subjects and strategically positioned objects simulating crowded scenarios, as shown in Figure 8. To diversify the scenarios, we arranged static objects in various configurations such as I-Configuration, L-Configuration, and T-Configuration setup, as exemplified in Figure 9. All setups were confined within an area of approximately 6 m $\times$ 10 m, with each data collection session covering a travel distance of 20 m and lasting approximately one minute. The human subjects in the setup were moving around during the data collection process.

Fig. 8. I-Shape path experimental set up which simulates a crowded indoor scene.

To ensure data quality, we captured all datasets in video format at a frame rate of 30 frames per second. Notably, meticulous attention was paid to environmental setup, with tables of dimensions approximately 80 cm $\times$ 60 cm $\times$ 250 cm (Height x Width x Length) arranged to emulate realistic scenarios.

A monocular camera mounted on the right side of the robot's platform of 1 m from the ground efficiently captured high-quality images and videos of our experimental setup, as illustrated in Figure 7(b). Video streams were input into DSO and results were recorded using the ROS rosbag functionality to generate corresponding (.bag) files containing exclusive DSO outputs. Point cloud coordinate information extracted from DSO outputs facilitated map construction in subsequent modules.

For map generation, we utilized the ROS map_server tool [34], enabling map creation and storage in image format (.png) for visualization or in Portable Gray Map format (.pgm) to facilitate evaluation tests. Before map evaluation and ablation studies, a ground truth reference was established using the original DSO point cloud map. Manual measurements of table sizes and annotation of ground-truth objects were performed, followed by post-processing steps including dilation [35], denoising [36], and C-obstacle [37] analysis to refine the map and minimize noise, ensuring accuracy and reliability.

B. Performance Index

The evaluation method adheres to the standard for map quality assessment, employing a journey-based approach [38] known for its reliability despite its relatively higher computational cost. The basic idea of [38] is to consider potential map users and determine map quality by averaging the quality of their actual path planning. More specifically, map users are characterized by journeys with clear starting and ending locations and the objective of navigating using shortest-path planning algorithms. The evaluation process involves comparing each map user's shortest path to the oracle's shortest path as determined by a manually annotated ground truth traversability map. For each waypoint along the oracle's path, the closest corresponding waypoint on the map user's path is determined, and the Euclidean distance between them is calculated as the error. These error values are averaged over all waypoints on the oracle's path and then

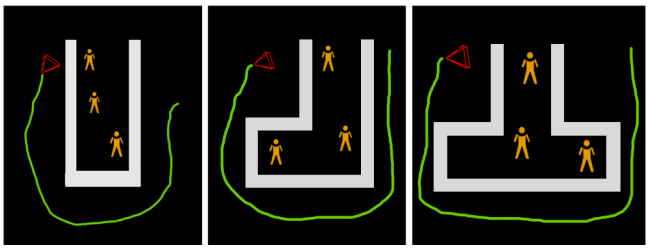


Fig. 9. Bird's eye view of obstacles setup of all kinds of configurations, namely I-Configuration, L-Configuration, and T-Configuration. The gray rectangle box indicates the point cloud data from DSO, hence the tables set up. The green trail is the frame positional data from DSO, and the red triangle is the current camera position, or current frame. Our data collection process is by using a robot equipped with a monocular camera and taking a video footage surrounding the set up.

averaged over all map users to obtain a map quality index.

The above journey-based metrics provide higher confidence and more realistic utility estimates compared to image processing-based approximations such as least squares error. In contrast, a potential concern with this metric is the increase in unbounded computational cost as the map size or number of potential map users increases. However, given the scale of our experiment, we found that the computation time was reasonable.

To enhance the thoroughness and accuracy of map evaluation, we developed specialized code capable of concurrently evaluating multiple maps. This code efficiently processes and analyzes numerous maps simultaneously, assigning priorities to each and generating an error score as output. This advanced evaluation methodology ensures the reliability and precision of our map evaluation process, facilitating a comprehensive assessment of overall system performance. Additionally, the simultaneous evaluation of multiple maps enables the identification of potential issues or discrepancies, empowering us to make necessary adjustments for optimal functionality.

C. Quantitative Evaluation

The existing method First-person view IMU (FPV IMU) is directly extended to camera TPV and is used as a baseline method. Furthermore, it is employed for direct comparisons between the proposed method and the best-known methods. In this context, we use shorter abbreviations for the modules to represent them for better readability: PHYS - Physical Constraint, PHOT - Photometric Constraint, W2M - Walk2Map++. By employing various combinations of modules, we can generate 7 distinct combinations, namely: (PHYS + PHOT + W2M, PHYS + PHOT, PHYS + W2M, PHOT + W2M, PHYS, PHOT, W2M). Each of these configurations produces its own performance score which we further evaluate qualitatively.

$$f_{\text{EvaluatedError}} = \frac{\sum_{i=1}^{n} |p_{\text{gt}}[i] - p_{\text{map}}[i]|}{N} \tag{3}$$

To comprehensively evaluate our system, we conducted independent tests for each combination. We meticulously recorded the performance metrics and conducted thorough

comparisons. Our experimentation encompassed three diverse datasets, and the resulting findings are presented in Table I. I-Cfg = I-Configuration; L-Cfg = L-Configuration; T-Cfg = T-Configuration. The average performance results are calculated with Equation 3.

From the findings presented in Table I, it is evident that the proposed method (PHYS + PHOT + W2M), consistently delivers robust performance across various configurations. While a lower average score indicates a better performing result, our method exhibits commendable efficacy even in more intricate scenarios with increasing complexity from I-configuration to L-configuration and T-configuration.

In comparing individual traversability maps (PHYS, PHOT, and W2M) with our proposed method, it is readily apparent that our approach outperforms them across all data configurations. Notably, the PHOT module utilizes an occlusion ordering algorithm to extrapolate the likely human presence within densely populated scenes, offering a generalized area without specifying precise coordinates. In contrast, the W2M module contributes to human distance estimation, translating human location from 3D to 2D space. The integration of W2M with the point cloud data from PHYS enables accurate human location estimation, ensuring a comprehensive and precise determination of the individual's position.

On a side note, it is important to acknowledge that each of the PHYS, PHOT, and W2M modules comes with its limitations and prerequisites for optimal functionality. For PHOT and W2M, challenges arose when the human upper body was not visible, fully occluded, or when multiple humans were standing together, leading to sub-optimal algorithm performance. In the case of PHYS, if DSO loses track of frames, the point cloud map output might fall short of expectations.

These modules collectively contributed to deducing the location of the human in 2D space. While the occlusion ordering algorithm provided an approximate location of the human, Walk2Map++ refined this estimation, offering a more accurate measurement of the human's distance from the camera. Each module generated its traversability map, and the final map was derived from the intersection of multiple maps, resulting in a more precise representation.

In instances where PHYS encountered difficulties, rendering the occlusion ordering algorithm unusable due to the absence of point clouds, our framework intelligently avoided attempting traversability map generation. Similarly, if PHOT faces issues, such as the inability to locate the human, it leads to the absence of a traversability map. In cases of Walk2Map++ failure, where the human distance cannot be determined accurately, a map was still generated, albeit as an estimation of the human area without precise human location information.

In situations where any of these modules confronted the aforementioned challenges, our framework adopted a judicious approach by refraining from generating the traversability map until the requisite conditions were met, ensuring the reliability and accuracy of the output.

TABLE I
AVERAGE PERFORMANCE RESULTS.

Average Performance	I-Cfg.	L-Cfg.	T-Cfg.
[39]	2.35	18.68	15.77
PHYS + PHOT + W2M	1.36	12.42	15.45
PHYS + PHOT	1.12	15.43	13.45
PHYS + W2M	3.36	13.22	9.07
PHOT + W2M	4.56	19.23	15.67
PHYS	2.31	18.67	23.45
PHOT	8.66	14.56	18.56
W2M	11.22	12.34	20.45

Furthermore, in our prior study [39] using ORB-SLAM3 [40] as the PHYS method instead of DSO [31], together with an occlusion ordering algorithm and a novel human height-depth estimation module, the performance was evaluated. This configuration, combined with an occlusion ordering algorithm and an innovative human height-depth estimation module, underwent performance evaluation. Significantly, our current method (PHYS + PHOT + W2M) surpasses this prior approach, underscoring substantial advancements in our research. DSO utilizes a direct visual odometry approach, minimizing photometric errors between consecutive frames, proving advantageous in scenes with less texture or repetitive patterns where feature-based methods like ORB-SLAM3 may encounter challenges. Furthermore, DSO's direct approach enhances robustness in dynamic scenes, where feature points may change rapidly, while feature-based methods such as ORB-SLAM3 might struggle due to reliance on distinctive features susceptible to occlusions or scene changes [41].

We improved our human-height depth estimation method, replacing the previous assumption-based approach that led to inaccuracies. We now use a human pose estimation method and a new distance estimation algorithm. By using two keypoints from the human torso (Figure 5), we significantly improve estimation reliability. This allows for more accurate calculation of torso length and human location, enhancing the traversability map's reliability.

V. CONCLUSIONS

In conclusion, this paper presents a novel integrated framework, offering inventive solutions for traversability prediction through multiple pedestrian observations. Leveraging a unique occlusion ordering algorithm and a human pose estimation distance estimator, our approach addresses the challenges of traversability estimation. By interlinking modules, each generating its traversability map, we achieve a comprehensive prediction of traversable areas within crowded scenes.

The novelty of our study lies in its distinctive approach, giving an overall good performance in terms of accuracy and the quality of the generated traversability map. This valuable research contributes a unique perspective to the field, offering an innovative approach to solving the traversability prediction problem.

REFERENCES

[1] M. Benrabah, E. Randriamiarintsoa, C. O. Mousse, J. Morceaux, R. Aufrère, and R. Chapuis, "Dual occupancy and knowledge maps

management for optimal traversability risk analysis," in *26th International Conference on Information Fusion, FUSION 2023*. IEEE, 2023, pp. 1–6.

[2] C. Sevastopoulos and S. Konstantopoulos, "A survey of traversability estimation for mobile robots," *IEEE Access*, vol. 10, pp. 96 331–96 347, 2022.

[3] R. O. Chavez-Garcia, J. Guzzi, L. M. Gambardella, and A. Giusti, "Learning ground traversability from simulations," *IEEE Robotics Autom. Lett.*, vol. 3, no. 3, pp. 1695–1702, 2018.

[4] P. Papadakis, "Terrain traversability analysis methods for unmanned ground vehicles: A survey," *Eng. Appl. Artif. Intell.*, vol. 26, no. 4, pp. 1373–1385, 2013.

[5] M. Wermelinger, P. Fankhauser, R. Diethelm, P. Krüsi, R. Siegwart, and M. Hutter, "Navigation planning for legged robots in challenging terrain," in *2016 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*. IEEE, 2016, pp. 1184–1189.

[6] Y. Pan, X. Xu, Y. Wang, X. Ding, and R. Xiong, "GPU accelerated real-time traversability mapping," in *2019 IEEE Int. Conf. on Robotics and Biomimetics, ROBIO 2019, Dali, China, December 6-8, 2019*. IEEE, 2019, pp. 734–740.

[7] S. Palazzo, D. C. Guastella, L. Cantelli, P. Spadaro, F. Rundo, G. Muscato, D. Giordano, and C. Spampinato, "Domain adaptation for outdoor robot traversability estimation from RGB data with safety-preserving loss," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS*. IEEE, 2020, pp. 10 014–10 021.

[8] B. Suger, B. Steder, and W. Burgard, "Traversability analysis for mobile robots in outdoor environments: A semi-supervised learning approach based on 3d-lidar data," in *IEEE IEEE Int. Conf. Robotics and Automation, ICRA*. IEEE, 2015, pp. 3941–3946.

[9] S. Martin, L. Murphy, and P. Corke, "Building large scale traversability maps using vehicle experience," in *Experimental Robotics - The 13th Int. Sym. on Experimental Robotics, ISER*, ser. Springer Tracts in Advanced Robotics, J. P. Desai, G. Dudek, O. Khatib, and V. Kumar, Eds., vol. 88. Springer, 2012, pp. 891–905.

[10] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak, "Planning to explore via self-supervised world models," in *Proceedings of the 37th Int. Conf. on Machine Learning, ICML, series = Proceedings of Machine Learning Research, volume = 119, pages = 8583–8592, publisher = PMLR, year = 2020,*.

[11] E. F. Morales, R. Murrieta-Cid, I. Becerra, and M. A. Esquivel-Basaldua, "A survey on deep learning and deep reinforcement learning in robotics with a tutorial on deep reinforcement learning," *Intell. Serv. Robotics*, vol. 14, no. 5, pp. 773–805, 2021.

[12] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017*. IEEE, 2017, pp. 4644–4651. [Online]. Available: https://doi.org/10.1109/ICRA.2017.7989540

[13] L. Tai, S. Li, and M. Liu, "Autonomous exploration of mobile robots through deep neural networks," *International Journal of Advanced Robotic Systems*, vol. 14, no. 4, p. 1729881417703571, 2017.

[14] T. Kucner, M. Magnusson, S. Mghames, L. Palmieri, F. Verdoja, C. Swaminathan, T. Krajník, E. Schaffernicht, N. Bellotto, M. Hanheide, and A. Lilienthal, "Survey of maps of dynamics for mobile robots," *The International Journal of Robotics Research*, vol. 42, no. 11, pp. 977–1006, Sept. 2023.

[15] A. Alempijevic, R. Fitch, and N. Kirchner, "Bootstrapping navigation and path planning using human positional traces," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 1242–1247.

[16] P. Papadakis and P. Rives, "Binding human spatial interactions with mapping for enhanced mobility in dynamic environments," *Autonomous Robots*, vol. 41, pp. 1047 – 1059, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:37073354

[17] C. Mura, R. Pajarola, K. Schindler, and N. Mitra, "Walk2map: Extracting floor plans from indoor walk trajectories," *Computer Graphics Forum*, vol. 40, pp. 375–388, 05 2021.

[18] M. Everett, J. Miller, and J. P. How, "Planning beyond the sensing horizon using a learned context," *CoRR*, vol. abs/1908.09171, 2019. [Online]. Available: http://arxiv.org/abs/1908.09171

[19] B. Bescós, J. M. Fácil, J. Civera, and J. Neira, "Dynslam: Tracking, mapping and inpainting in dynamic scenes," *CoRR*, vol. abs/1806.05620, 2018. [Online]. Available: http://arxiv.org/abs/1806.05620

[20] R. Schmid, D. Atha, F. Scholler, S. Dey, S. Fakoorian, K. Otsu, B. Ridge, M. Bjelonic, L. Wellhausen, M. Hutter, and A.-a. Aghamohammadi, "Self-supervised traversability prediction by learning to reconstruct safe terrain," 10 2022, pp. 12 419–12 425.

[21] M. A. Saucedo, A. Patel, C. Kanellakis, and G. Nikolakopoulos, "Eat: Environment agnostic traversability for reactive navigation," *Expert Systems with Applications*, vol. 244, p. 122919, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423034218

[22] E. Chen, C. Ho, M. Maulimov, C. Wang, and S. Scherer, "Learning-on-the-drive: Self-supervised adaptation of visual offroad traversability models," 2023.

[23] M. V. Gasparino, A. N. Sivakumar, Y. Liu, A. E. B. Velasquez, V. A. H. Higuti, J. Rogers, H. Tran, and G. Chowdhary, "Wayfast: Navigation with predictive traversability in the field," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 651–10 658, 2022.

[24] J. Zhu, H. Zhou, Z. Wang, and S. Yang, "Improved multi-sensor fusion positioning system based on gnss/lidar/vision/imu with semi-tight coupling and graph optimization in GNSS challenging environments," *IEEE Access*, vol. 11, pp. 95 711–95 723, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3311359

[25] M. V. Gasparino, A. N. V. Sivakumar, and G. Chowdhary, "Wayfaster: a self-supervised traversability prediction for increased navigation awareness," *ArXiv*, vol. abs/2402.00683, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267364840

[26] S. Li, P. Kou, M. Ma, H. Yang, S. Huang, and Z. Yang, "Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data," *IEEE Access*, vol. 12, pp. 27 331–27 343, 2024.

[27] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. Springer, 2016, vol. 9.

[28] A. Monszpart, P. Guerrero, D. Ceylan, E. Yumer, and N. J. Mitra, "imapper: interaction-guided scene mapping from monocular videos," *ACM Transactions on Graphics*, vol. 38, no. 4, p. 1–15, July 2019. [Online]. Available: http://dx.doi.org/10.1145/3306346.3322961

[29] G. Appenzeller, J.-H. Lee, and H. Hashimoto, "Building topological maps by looking at people: an example of cooperation between intelligent spaces and robots," in *Proceedings of the 1997 IEEE/RSJ International Conference on Intelligent Robot and Systems. Innovative Robotics for Real-World Applications. IROS '97*, vol. 3, 1997, pp. 1326–1333 vol.3.

[30] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[31] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," 2016.

[32] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[33] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022.

[34] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org

[35] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[37] J.-C. Latombe, *Obstacles in Configuration Space*. Boston, MA: Springer US, 1991, pp. 105–152.

[38] D. C. Lee, *The Map-Building and Exploration Strategies of a Simple Sonar-Equipped Mobile Robot: An Experimental, Quantitative Evaluation*, ser. Distinguished Dissertations in Computer Science. Cambridge University Press, 1996.

[39] J. T. Y. Liang and K. Tanaka, "Walking = traversable? : Traversability prediction via multiple human object tracking under occlusion," 2023.

[40] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: an accurate open-source library for visual, visual-inertial and multi-map SLAM," *CoRR*, vol. abs/2007.11898, 2020. [Online]. Available: https://arxiv.org/abs/2007.11898

[41] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, "A comprehensive survey of visual slam algorithms," *Robotics*, vol. 11, no. 1, 2022.