Human-Centric Robot Navigation: Leveraging Pedestrian Occlusion Patterns for Traversability Analysis in Crowded Indoor Environments

Jonathan Tay Yu Liang Kanji Tanaka

***Abstract*— In the realm of Human-Robot Interaction (HRI), enabling robots to navigate crowded indoor spaces remains a significant challenge. While existing methods like Walk2Map show promise, they often rely on intrusive equipment, limiting their practicality in dynamic social settings. Our research takes a novel approach by analyzing human-environment interactions from a third-person perspective, eliminating the need for wearable sensors. By studying the interplay between pedestrians and static objects, we aim to develop a more intuitive system for robots to understand and navigate human-centric spaces. This method enhances robots' ability to respect human movement patterns and social dynamics and promotes seamless integration into crowded environments. We aim to advance socially aware robot navigation, improving human-robot coexistence and collaboration in various public spaces.**

I. Introduction

Traversability prediction through visual recognition of navigable areas on a 2D floor remains a fundamental challenge in robot navigation. While region-wise traversability prediction techniques [1]–[13] have been extensively studied for outdoor environments such as woods and streets, they face significant limitations in crowded, occlusion-prone indoor spaces like offices or classrooms, where human traffic and dynamic obstacles are prevalent. Recent studies have begun to address this gap by focusing on indoor environments. Kucner et al. [14] constructed maps of dynamics (MoDs) using laser range finders to encode semantic information on motion patterns. Alempijevic et al. [15] leveraged human interactions to map human motion dynamics, identifying areas of changing traversability. Papadakis et al. [16] proposed a generative methodology for indoor robot navigation that incorporated human spatial activity for passage detection and occupancy prediction, while mitigating false positives using prior map information. However, these approaches are primarily effective in scenarios with minimal obstacles and occlusions, leaving the challenges posed by typical office environments largely unaddressed.

In the realm of architectural design and map creation, a novel data-driven method called Walk2Map [17] has recently gained attention. This approach, which utilizes a first-person IMU to generate traversability maps, offers a simple yet powerful means of creating floor plans based solely on indoor pedestrian trajectories. Inspired by advancements in affordable, high-performance equipment such as smartphone IMUs and data-driven inertial odometry, Walk2Map produces floor plans of exceptional quality that exhibit ideal characteristics for use as traversability maps [18] in mobile robotics. However, the method's application in autonomous mobile robot navigation is severely limited by the need to equip pedestrians with odometers, as illustrated in Figure 1.

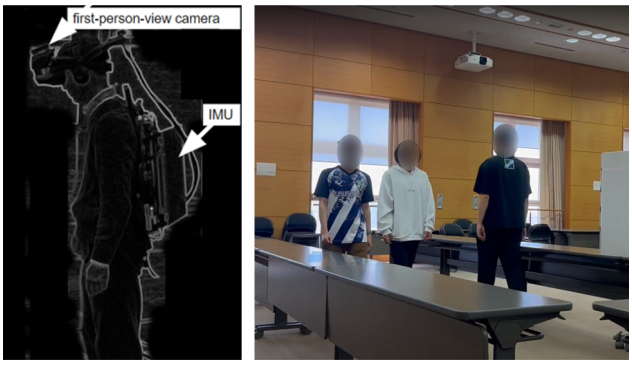


Fig. 1. Traversability prediction under severe occlusion. Left: Conventional first-person-view setup with IMU. Right: Proposed third-person-view monocular vision setup.

Inspired by the remarkable outcomes observed, we propose Walk2Map++, an expansion of Walk2Map that incorporates vision-based methodologies to enhance Human-Robot Interaction (HRI) in indoor environments. While implementing a third-person robot vision setup presents a notable challenge due to its inherently ambiguous nature compared to first-person IMU setups, we anticipate that recent advancements in deep learning-aided third-person human behavior analysis techniques [19] are sufficiently mature to provide effective visual measurements for improved robot-human coexistence. Diverging from the conventional approach of eliminating dynamic information from the scene, we propose leveraging these dynamic cues to enhance our exploration of traversable regions, allowing robots to navigate more naturally in human-populated spaces. We demonstrate that by delving deeper into the intricate physical and photometric interactions between humans and obstacles, we can derive enhanced cues that facilitate the reconstruction of pedestrian trajectories, leading to more socially aware robot navigation. To address the ill-posed nature of the problem and improve HRI, we propose leveraging photometric and physical cues, specifically focusing on human-object occlusion reasoning and collision avoidance. These cues are inferred from past and present observations, aided by object reconstruction and human tracking techniques, ultimately enabling robots to better understand and respond to human behavior in shared spaces.

This paper presents several key contributions to enhance

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

Human-Robot Interaction (HRI) through improved robot navigation in shared spaces: (i) We introduce Walk2Map++, an innovative approach that transforms Walk2Map's first-person IMU sensor into a third-person view from a robot's onboard camera, allowing for more natural and unobtrusive observation of human behavior. (ii) We establish dynamic relationships and physical cues between moving humans and stationary objects using SLAM and human detectors, enabling a deeper understanding of human-environment interactions. (iii) We present a novel third-person view traversability estimation approach that combines SLAM and human detection methods, replacing the first-person view IMU method and allowing robots to interpret human movement patterns for more socially aware navigation. (iv) We evaluate our method's effectiveness through a performance index for traversability maps, validated via fusion and comparison with well-known methods in comprehensive real-world experiments, demonstrating its potential for improved robot navigation in human-centric environments. (v) To foster further research and adoption in HRI, we will make the code and datasets associated with this study publicly available upon acceptance. These contributions collectively aim to significantly enhance robots' ability to understand, predict, and adapt to human behavior in shared spaces, ultimately leading to more harmonious human-robot coexistence in various indoor environments.

II. Related Works

A. Traversability Prediction

Traversability prediction, a prominent field of computer vision research, has undergone notable advancements within the supervised learning paradigm [18]. 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 [20]. A noteworthy trend involves the exploration of self-supervised frameworks for autonomous robot applications, addressing challenges such as long-range traversability [21], RGB-D traversability prediction [22], visibility challenging environments [23], and hazardous forest scenarios [24]. In addition to supervised learning, emerging semi-supervised [25] and unsupervised [26] frameworks which leverage 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 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 [27], a data-driven method that uniquely leverages the correlation between human-object interactions and scene-object arrangements. By identifying characteristic interactions and employing an 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 [28]. 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 shows 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 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 1. 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, by leveraging the physical cue (PHYS) 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 reasoning between humans and obstacles, a photometric cue (PHOT) 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 optimiza-

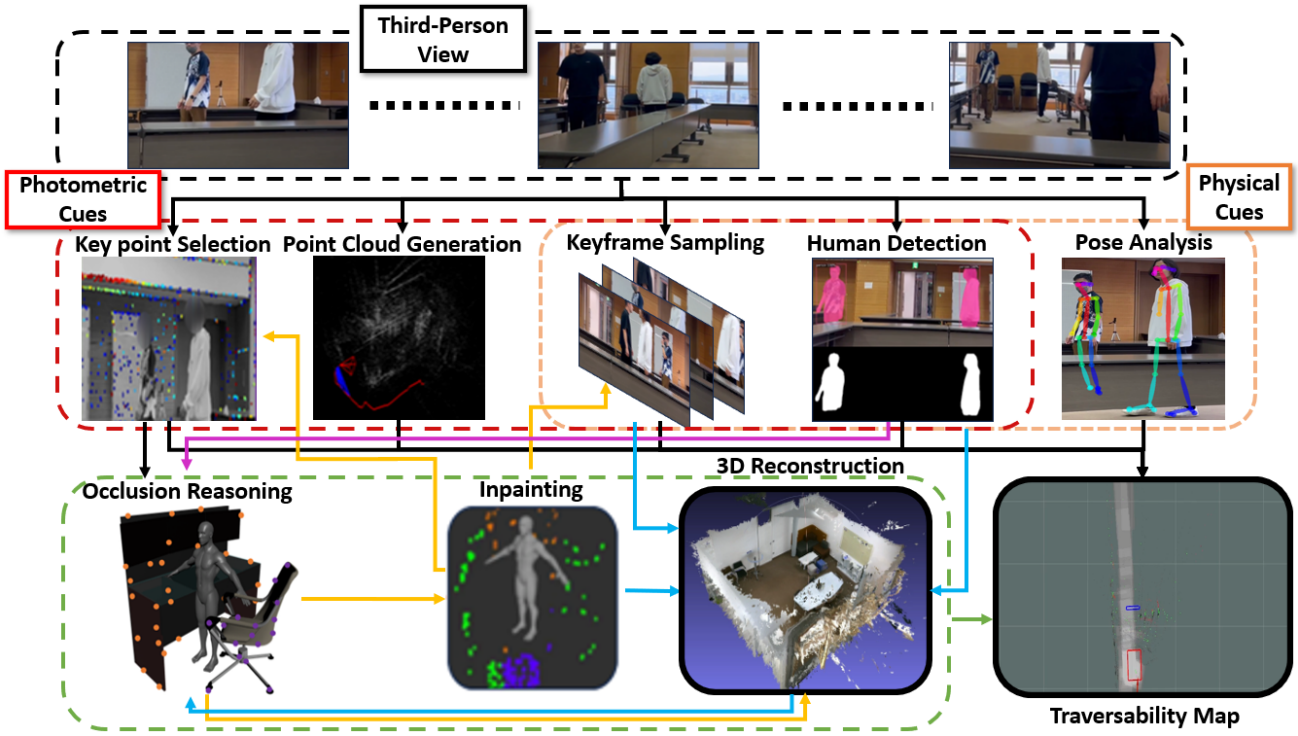


Fig. 2. Block diagram of framework: All modules are interconnected via ROS (Robot Operating System). The PHYS module comprises DSO (Direct Sparse Odometry) and is responsible for generating point clouds utilized by other modules. Within the PHOT module, a Human-Object Occlusion Ordering Algorithm is 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. The red box represents the estimated human location in the traversability map image, hence the traversable region. The grey path indicates the traversable region, in which the human has walked. The grey path is imprinted by the red boxes.

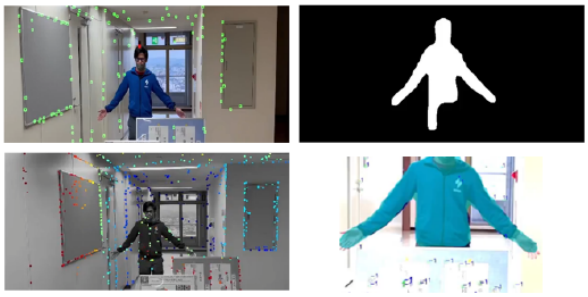


Fig. 3. Top-left: The projection of point clouds to the keyframes visualize; Top-right: Human mask used for occlusion ordering algorithm; Bottom-left: DSO; Bottom-right: Detectron2 Semantic Segmentation

tion during loop closing (Section III-E). Figure 3 shows the different modules of our framework.

A. Traversability Map

The traversability map constitutes a vital component within the framework of the robotic system, manifesting as a two-dimensional grid overlaying the mobile plane of the robot. The grid cell 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” - navigable terrain; “untraversable” - impassable regions; and “unknown” - 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.

B. Walk2Map++

To estimate the human distance from the camera, deep learning, and semantic segmentation methods are typically used. In [29], the proposed pipeline utilizes MobileNetV1 [30] as the backbone network, together with Atrous Spatial Pyramid Pooling (ASPP) [31] to construct the encoder. Despite its strong performance, the method faces challenges when dealing with variations in in-depth data across different parts of the human body, especially in scenarios where parts of the body are occluded, such as when handling large objects like boxes or furniture. To address this, we propose leveraging the entire body of the pedestrian, considering the co-occurrence of the head, arms, torso, and legs to reduce detection errors. Specifically, we employ a human pose estimator to identify the key points of the human torso and then use a pinhole camera model to infer the human’s distance from the camera.

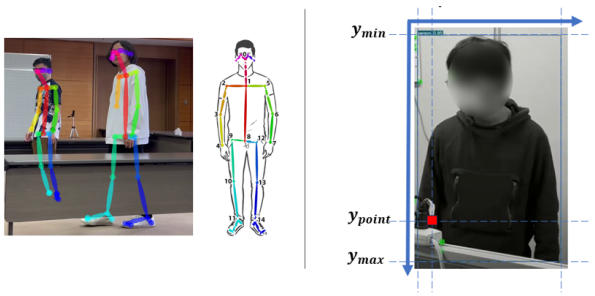


Fig. 4. (Left, Middle): Human-centric coordinate system. As shown in the middle figure, keypoint 1 and keypoint 8 is used as a reference point of torso length.(Right): The relationship between occluder's feature point and occluded human's region.

To distinguish between pedestrians and non-pedestrians, we use the pedestrian's entire body as an error detection code. On the other hand, 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.

$$D = k \cdot \frac{f \cdot L}{\|P_1 - P_2\|} \tag{1}$$

Specifically, the measured torso length is denoted as L and the distance from the camera to the human torso is 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 1, where Euclidean distance calculation is employed. In this study, we utilize the state-of-the-art human pose estimation model - OpenPose [32], for this purpose.

Before model implementation, parameter calibration is essential, with a focus on determining the key parameter k in Equation 1. Utilizing two keypoints from human pose estimation, such as those illustrated in Figure 4, 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. Instead, 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.

C. Physical Cues (PHYS)

Previous algorithms like UnionDet [33] and PPDM [34] use parallel HOI (Human-Object Interaction) detectors based on interaction or union boxes, often requiring heuristic thresholding for post-processing. In this study, we integrate PHYS and PHOT to efficiently establish human-object relationships. By continuously updating point cloud coordinates and human location data, our approach provides real-time estimates of human location, enhancing understanding of human-object interactions in dynamic scenes.

Fig. 5. 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.

For this purpose, we employ DSO [35] because this specific SLAM algorithm provides high discrimination ability between dynamic and static objects, yielding a reasonably dense point cloud format obstacle map, as shown in Figure 5. Moreover, it is necessary to note that obstacles positioned higher than human height do not pose physical cues to 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 the 2D space of the 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 Walk2Map++ and the occlusion reasoning algorithm is found. The intersection between traversability maps is updated into the final map.

D. Photometric Cues (PHOT)

Occlusion reasoning poses a persistent challenge in computer vision, with various approaches proposed to tackle it. Sun et al. introduced a method using layered image motion with explicit occlusions for depth ordering [36], yet dynamic scenes with complex geometries and occlusions remain challenging. Our algorithm, employing both PHYS and PHOT modules, effectively addresses this challenge. Leveraging DSO's point cloud data for precise static object localization and iterative processing for establishing occlusion cues, we continuously track human subjects and dynamically update their locations. By projecting point clouds onto keyframes and integrating human region data from Detectron2 [37], our algorithm accurately determines human location, while also handling scenarios with overlapping human instances by deferring processing until only one human is detected.

Figure 4 displays a red point at coordinates (x_{point} and y_{point}), within the human area defined by (x_{min}, x_{max}, y_{min}, and y_{max}). Pedestrians are typically found among clusters of point clouds, suggesting higher traversability probabilities in these regions. if x_{point} and y_{point} fall within the human area, the point is likely in front of the human.

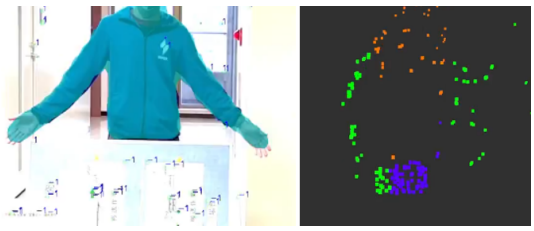
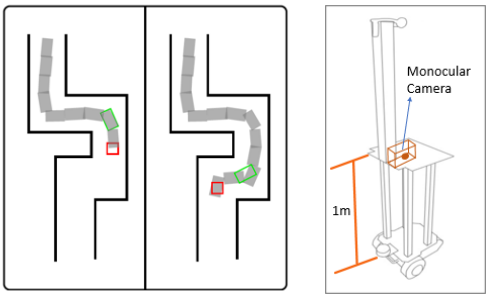


Fig. 6. Left: Occlusion ordering algorithm; Right: Grouped cluster point cloud visualized in rviz visualizer. Purple cluster indicate points that are in front of the human, orange cluster indicate points that are behind of the human.



(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

The grid map assigns each cell one of these three states: "traversable", "untraversable", or "unknown". These values in the grid map are adjusted based on point clusters, aiding in human location inference.

When the human body is fully occluded or when most of the body is occluded, the algorithm will wait until the upper body of the human becomes fully visible or when the human tracker can track the human again.

E. Asynchronous Map Fusion

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 [35], Human pose estimation for human distance estimation, and human occlusion ordering algorithm. As shown in Equation 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 cues using a novel human-object occlusion ordering algorithm (Section III-D).

$$T_{\text{combined}} = T_1 \cap T_2 \tag{2}$$

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.

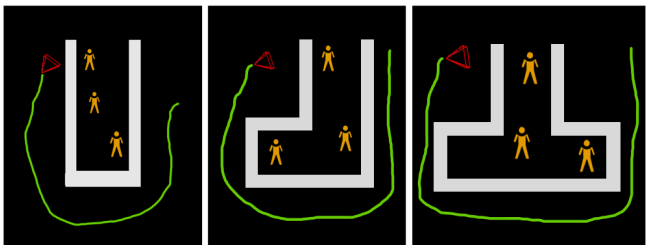


Fig. 8. Bird's eye view of obstacles setup of all kinds of configurations, namely I-Configuration, L-Configuration, and T-Configuration. The gray rectangle box indicates the point cloud data from DSO, hence the tables set up. The green trail is the frame positional data from DSO, and the red triangle 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.

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 [37] human mask and YOLOv7 [38] 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, as shown in Figure 6. 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 collect three distinct datasets, each comprising multiple human subjects and objects simulating crowded scenarios. To diversify the scenarios, we arranged static objects in various configurations such as I-Configuration, L-Configuration, and T-Configuration setup, as shown in Figure 8. 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.

To ensure data quality, we captured all datasets in video format at a frame rate of 30 frames per second. Tables of dimensions approximately 80 cm $\times$ 60 cm $\times$ 250 cm (Height x Width x Length) were arranged in the experimental setup to emulate realistic scenarios, as shown in Figure 9. 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

Fig. 9. I-Shape path experimental set up which simulates a crowded indoor scene.

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 [39], enabling map creation to facilitate evaluation tests. Before map evaluation and ablation studies, a ground truth reference was established by manually annotating 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 [40], denoising [41], and C-obstacle [42] analysis to refine the map and minimize noise, ensuring accuracy and reliability.

B. Performance Index

We employ a journey-based approach [43] for map quality assessment, a method recognized for its reliability despite higher computational requirements. This technique evaluates map quality by simulating realistic path-planning scenarios. User journeys, defined by distinct start and end points, are analyzed using shortest-path algorithms. We quantify errors by comparing each simulated user path to an oracle path derived from a manually annotated ground truth map, calculating distances between corresponding waypoints. The average error is then computed across a large sample of users. Our custom-developed code optimizes evaluation efficiency through concurrent analysis of multiple maps, priority assignment, and generation of error scores. This comprehensive methodology ensures thorough and accurate map assessment, facilitating robust system performance evaluation and pinpointing areas for improvement.

The journey-based metrics offer superior confidence and more realistic utility estimates compared to image processing-based approximations such as least squares error. While a potential drawback of this approach is the unbounded computational cost as map size or user numbers increase, we found the computation time reasonable within the scale of our experiments. To further enhance evaluation thoroughness and accuracy, we developed specialized software capable of concurrent multi-map assessment. This advanced tool efficiently processes and analyzes numerous maps simultaneously, assigning priorities and generating error scores for each. Our sophisticated evaluation methodology ensures reliable and precise map assessment, enabling comprehensive system performance analysis. Moreover, the parallel evaluation of multiple maps facilitates the rapid identification of potential issues or discrepancies, allowing for timely adjustments and optimization of system functionality.

C. Quantitative Evaluation

The existing First-person view IMU method is directly extended to the camera's Third-person view 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 better readability: PHYS - Physical Cues, PHOT - Photometric Cues, 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 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. In this context, a lower average score indicates a better-performing result. The average performance results are calculated with Equation 3.

From the findings presented in Table I, it shows that the proposed method (PHYS + PHOT + W2M), consistently delivers robust performance across various configurations. Our method exhibits commendable efficacy even in more intricate scenarios with increasing complexity from I-configuration to L-configuration and T-configuration.

Our method surpasses individual traversability maps (PHYS, PHOT, and W2M) across various data setups. While PHOT employs an occlusion reasoning algorithm to estimate human position, W2M focuses on translating human 3D location to 2D space for accurate distance estimation. Combining W2M with PHYS's point cloud data enhances human location precision in 2D. Despite challenges like absent point clouds in PHYS or human location failure in PHOT, our framework intelligently refrains from generating traversability maps. In instances of Walk2Map++ malfunction, a map is still produced, though with estimated human area. Our method ensures reliability and accuracy by withholding traversability map generation until necessary conditions are met. By utilizing two keypoints from the human torso as shown in Figure3(a), we've enhanced the reliability of our estimations. This eliminates the inaccuracies caused by assuming a fixed height range for humans and provides a more confident calculation of torso length, leading to improved accuracy in determining human location.

TABLE I

AVERAGE PERFORMANCE RESULTS.

Average Performance	I-Cfg.	L-Cfg.	T-Cfg.
[44]	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

V. CONCLUSIONS

In conclusion, this paper presents a solution that addresses the challenges of traversability prediction in dynamic, human-populated environments, with a focus on enhancing Human-Robot Interaction (HRI). By leveraging an occlusion reasoning algorithm and a human pose estimation distance estimator, our approach extends beyond traditional vSLAM methods to account for the presence and behavior of humans in the environment. Through the integration of physical and photometric cues, we generate a traversability map that considers human dynamics and spatial relationships.

Our study offers a distinctive approach to HRI by establishing a relationship between dynamic humans and static objects to predict traversable regions in indoor scenes. This method enables robots to navigate crowded spaces more effectively, leading to smoother and more natural interactions with humans. The comprehensive prediction of traversable areas within human-occupied scenes aligns with the need for richer environmental understanding in HRI scenarios, especially when limited observations are available.

The promising performance in terms of accuracy and the quality of the generated traversability map demonstrates the potential for improving robot navigation in human-centric environments. This research contributes a unique perspective to the field of HRI, offering an innovative approach to solving the traversability prediction problem while considering human presence and movement patterns. By enhancing robots' ability to understand and navigate human-populated spaces, our work paves the way for more seamless and intuitive human-robot coexistence in various indoor settings.

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, Charleston, SC, USA, June 27-30, 2023*. IEEE, 2023, pp. 1–6. [Online]. Available: https://doi.org/10.23919/FUSION52260.2023.10224224

[2] C. Sevastopoulos and S. Konstantopoulos, "A survey of traversability estimation for mobile robots," *IEEE Access*, vol. 10, pp. 96 331–96 347, 2022. [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3202545

[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. [Online]. Available: https://doi.org/10.1109/LRA.2018.2801794

[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. [Online]. Available: https://doi.org/10.1016/j.engappai.2013.01.006

[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 International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*. IEEE, 2016, pp. 1184–1189. [Online]. Available: https://doi.org/10.1109/IROS.2016.7759199

[6] Y. Pan, X. Xu, Y. Wang, X. Ding, and R. Xiong, "GPU accelerated real-time traversability mapping," in *2019 IEEE International Conference on Robotics and Biomimetics, ROBIO 2019, Dali, China, December 6-8, 2019*. IEEE, 2019, pp. 734–740. [Online]. Available: https://doi.org/10.1109/ROBIO49542.2019.8961816

[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 International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021*. IEEE, 2020, pp. 10 014–10 021. [Online]. Available: https://doi.org/10.1109/IROS45743.2020.9341044

[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 International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*. IEEE, 2015, pp. 3941–3946. [Online]. Available: https://doi.org/10.1109/ICRA.2015.7139749

[9] S. Martin, L. Murphy, and P. Corke, "Building large scale traversability maps using vehicle experience," in *Experimental Robotics - The 13th International Symposium on Experimental Robotics, ISER 2012, June 18-21, 2012, Québec City, Canada*, ser. Springer Tracts in Advanced Robotics, J. P. Desai, G. Dudek, O. Khatib, and V. Kumar, Eds., vol. 88. Springer, 2012, pp. 891–905. [Online]. Available: https://doi.org/10.1007/978-3-319-00065-7_59

[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 International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 8583–8592. [Online]. Available: http://proceedings.mlr.press/v119/sekar20a.html

[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. [Online]. Available: https://doi.org/10.1007/s11370-021-00398-z

[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, Sep. 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] 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.

[19] N. Hirose, A. Sadeghian, M. Vazquez, P. Goebel, and S. Savarese, "Gonet: A semi-supervised deep learning approach for traversability estimation," 10 2018, pp. 3044–3051.

[20] 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

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

[22] 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.

[23] 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

[24] 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

[25] 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. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3367772

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

[27] 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, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1145/3306346.3322961

[28] G. Appenzeller, J.-H. Lee, and H. Hashimoto, "Building topological maps by looking at people: an example of cooperation between intelligent spaces and robots," 10 1997, pp. 1326 – 1333 vol.3.

[29] S. An, F. Zhou, M. Yang, H. Zhu, C. Fu, and K. A. Tsintotas, "Real-time monocular human depth estimation and segmentation on embedded systems," *CoRR*, vol. abs/2108.10506, 2021. [Online]. Available: https://arxiv.org/abs/2108.10506

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

[31] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: http://arxiv.org/abs/1802.02611

[32] 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.

[33] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," 2023.

[34] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," 2020.

[35] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *CoRR*, vol. abs/1607.02565, 2016. [Online]. Available: http://arxiv.org/abs/1607.02565

[36] D. Sun, E. Sudderth, and M. Black, "Layered image motion with explicit occlusions, temporal consistency, and depth ordering," in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010. [Online]. Available: https://proceedings.neurips.cc/paper$_files/paper/2010/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf$

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

[38] 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.

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

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

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

[42] J.-C. Latombe, *Obstacles in Configuration Space*. Boston, MA: Springer US, 1991, pp. 105–152. [Online]. Available: https://doi.org/10.1007/978-1-4615-4022-9$_3$

[43] A. M. Andrew, "The map-building and exploration strategies of a simple sonar-equipped mobile robot: An experimental quantitative evaluation, by david lee, distinguished dissertations in computer science series, cambridge university press, cambridge, 1996, xi+228 pp., isbn 0-521-57331-9 (hbk: £35)," *Robotica*, vol. 15, no. 2, p. 233–236, 1997.

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