CoarseBIMLoc: BIM-Integrated Coarse Global Localization via Open-Set Scene Graphs

Chang Chen[1,5], Junyou Chen[2], Yinqiang Zhang[1], Ruihua Han[1], Ming Zhang[3], Liang Lu[4], Jia Pan[1,5,†]

***Abstract*—We focus on coarse global localization in GPS-denied indoor environments by leveraging semantic priors from Building Information Modeling (BIM). Existing BIM-based methods often depend on synthetic image rendering, closed-set object detection for initialization, or dense point-cloud registration, and are limited by texture-poor visual ambiguity, weak open-set generalization, and high computational cost in large buildings. We propose *CoarseBIMLoc*, a BIM-integrated framework that combines an offline hierarchical BIM graph with an online sparse open-set scene graph. Offline, the BIM model is converted into a floor-room-object hierarchical graph with semantic labels and room topology. Online, vision-language models incrementally construct a 3D scene graph from RGB-D observations through cross-frame instance association and fusion. Localization is formulated as a semantic Monte Carlo localization problem, in which multiple pose hypotheses are maintained and particle weights are updated using both semantic and spatial correspondences between online observations and BIM objects. We evaluate the method in Gazebo on a BIM floor derived from a large real-world building in Hong Kong using a Clearpath Jackal platform. Across six trajectories, *CoarseBIMLoc* achieves reliable relocalization with approximately $0.6\,\mathrm{m}$ translation error, 12° orientation error, and 92.7% room-level localization accuracy, providing a robust coarse pose prior for downstream fine-grained registration and mapless navigation.**

***Index Terms*—Building Information Modeling, scene understanding, open-set scene graphs, Monte Carlo localization.**

I. Introduction

As digital workflows become standard in building design, construction, and operation, Building Information Modeling (BIM) has emerged as a foundational representation across the building lifecycle. Through the Industry Foundation Classes (IFC) schema, BIM provides a structured digital twin that captures detailed geometry, material attributes, and topological relationships. Beyond traditional use cases such as risk mitigation and construction monitoring, BIM is increasingly explored as a prior for GPS-denied indoor localization and robotic autonomy [1]–[8]. Compared with conventional priors such as static floor plans and 2D venue maps [9], BIM offers unique spatial intelligence, semantic information, large-scale scene coverage, and multi-floor structure. These properties make BIM a compelling foundation for *mapless navigation*, where robots can avoid pre-building dense, SLAM-based high-definition (HD) maps for every deployment site.

Despite this promise, translating static BIM assets into robust online global localization remains difficult. Traditional methods primarily rely on dense geometric features from BIM models for registration. Li et al. [4] rendered images from BIM for retrieval and proposed a domain-adaptive approach to bridge the gap between synthetic and real images. Zhang et al. [8] used dense point clouds from BIM for global registration with online sensor data. Nevertheless, these approaches are still limited by the visual ambiguity of low-level image features and high computational overhead, which impedes real-time deployment on edge platforms.

Semantic context, i.e., object categories and their spatial arrangements, offers a robust complementary signal to address these geometric and visual ambiguities. Such high-level cues are often more distinctive than low-level texture and can substantially improve coarse global localization. Although recent studies have shown the value of semantic priors for offline point-cloud classification [5], robot initialization [6], and automated 3D scanning [7], their integration into active online BIM-grounded localization in open-world scenarios remains underexplored.

To address this gap, we propose *CoarseBIMLoc*, a semantic-aware framework that treats BIM as a hierarchical scene graph rather than a static geometric mesh. We encode BIM object categories into CLIP text embeddings and match them with online instance embeddings extracted from robot observations, without requiring photorealistic image rendering. Rather than estimating a single global transform [9], our particle-based formulation maintains multimodal pose uncertainty to handle visual ambiguity. This design bridges structured architectural priors and online robotic perception through three core components:

1) **Offline Hierarchical BIM-Graph Construction.** We extract a multilevel (floor-room-object) scene graph from IFC data and introduce a language-encoded room classification strategy, capturing both semantic hierarchy and site topology.
2) **Online Open-Set Scene Graph Construction.** Unlike traditional CNN-based methods [6], we leverage recent vision-language models (VLMs) [10], [11] and 2D-to-3D fusion [12] to build an open-set scene graph in real time, improving robustness under partial observations and multiview inconsistency.

[1]Chang Chen, Yinqiang Zhang, Ruihua Han, and Jia Pan are with the School of Computing and Data Science, The University of Hong Kong, Pok Fu Lam, Hong Kong, China. (changchen@connect.hku.hk, zyq507@connect.hku.hk, hanrh@connect.hku.hk, jpan@cs.hku.hk)

[2]Junyou Chen is with the School of Control Science and Engineering, Shandong University. This work was done during an internship. (JunyouChen@mail.sdu.edu.cn)

[3]Ming Zhang is with the Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China. (mzhang457-c@my.cityu.edu.hk)

[4]Liang Lu is with the School of Artificial Intelligence, Shenzhen University, China. (llu@szu.edu.cn)

[5]Chang Chen and Jia Pan are also with the Shenzhen Loop Area Institute (Corresponding author: Jia Pan).

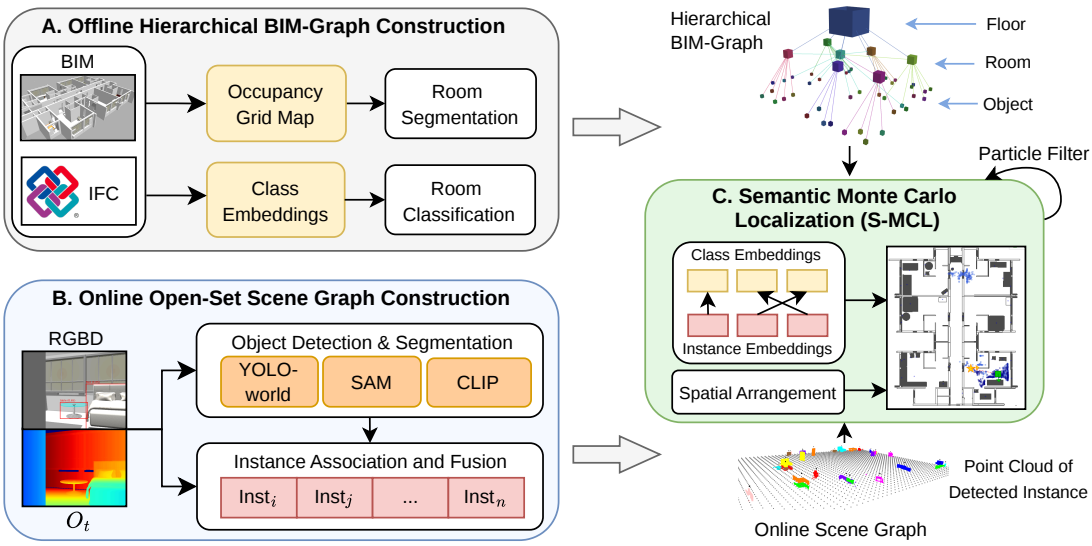

Fig. 1. Overview of the proposed *CoarseBIMLoc* framework. **Offline stage:** Architectural BIM (IFC) assets are parsed to construct a hierarchical floor-room-object prior scene graph with rich semantic labels and topological connectivity. **Online stage:** Live RGB-D streams are processed by open-set vision-language models to incrementally build a 3D open-set scene graph. **Localization stage:** Semantic and spatial correspondences between the online observations and the offline BIM graph nodes form measurement likelihoods for particle filter reweighting. This process yields a robust coarse global pose estimate, thereby significantly narrowing the search space for downstream fine-grained geometric registration and navigation.

3) **Semantic Monte Carlo Localization (S-MCL).** We introduce a semantic particle-filter formulation that iteratively estimates global pose by jointly evaluating category semantics and spatial arrangement consistency in a polar-coordinate view, which effectively handles object ambiguity.

We evaluate the method in Gazebo simulation using a BIM model derived from a large real-world building in Hong Kong. Results show that our algorithm achieves approximately $0.6\,\mathrm{m}$ translation error, 12° orientation error, and 92.7% room-level localization accuracy, yielding a reliable prior for fine-grained geometric registration and mapless navigation tasks.

II. Methodology

Our objective is to estimate the global pose $g_t = (x_t, y_t, \theta_t)$ in the BIM coordinate frame from RGB-D streams using a semantic Monte Carlo localization framework. We model the BIM environment as a bounded volume $\mathbb{V} \subset \mathbb{R}^3$ containing multiple room regions. Following open-set scene-graph principles [12], [13], we represent scene structure and semantics as a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where vertices $v_i \in \mathcal{V}$ encode object-level semantic embeddings and edges $e_{ij} \in \mathcal{E}$ denote spatial relations. In the offline stage, we construct a BIM-grounded hierarchical graph $\mathcal{G}$ from IFC data. In the online stage, the robot acquires RGB-D observations $O_t = (I_t, D_t)$ at timestep t and incrementally builds a partial open-set scene graph memory $\mathcal{M}$ using camera intrinsics K and odometry estimates p_t. In parallel, the localization module aligns online detected instances with the offline BIM graph and reweights particles with semantic and spatial evidence whenever new observations arrive. This process repeats until the particle states converge.

A. Offline Hierarchical BIM-Graph Construction

Given a BIM model of the scene, we first generate a localization-oriented offline prior composed of an occupancy grid map, structured room/object metadata, navigable free space, and a hierarchical scene graph. We load the BIM model into Gazebo and obtain the occupancy grid map (OGM) through ray-casting simulation. We then segment rooms via connected-component analysis and a flood-fill algorithm, while door elements are used to infer room connectivity. Object-to-room assignment is performed by mapping object coordinates to room masks. To assign room categories, we propose an efficient language-encoded room classification method, which concatenates object labels inside each room into a sentence (e.g., "sink, toilet, bathtub"), encodes it with a lightweight sentence encoder [14], and matches it against a predefined room-category embedding pool to obtain the best-matching room type. The final offline representation is a floor-room-object hierarchical scene graph $\mathcal{G}$ with CLIP text embeddings C_k on object nodes and room and object connectivity on edges.

B. Online Open-Set Scene Graph Construction

To emulate human navigation using sparse semantic cues and topology, we maintain an incremental open-set scene graph representation $\mathcal{M}$. Given RGB-D observations $O_t = (I_t, D_t)$ at timestep t, we use YOLO-world open-set detector [15] with a household category vocabulary to obtain 2D region proposals R_t^{2D}. Each proposal crop is encoded by CLIP [10] (ViT-L) to produce a visual embedding f_j^{2D} and refined into a binary mask by Segment Anything (SAM) [11]. Using the depth map D_t and camera intrinsics K, the masked regions are lifted to 3D instances R_t^{3D} and denoised using DBSCAN clustering. To handle partial observations and multiview inconsistency, detections are associated across frames. Instances from different timesteps are transformed to the world frame using the robot pose p_t and merged when both cosine similarity and volumetric overlap IoU_3 exceed predefined thresholds δ_{sim} and δ_{IoU}, otherwise a new node is initialized. If matched, we perform both volumetric and embedding fusion:

$f_k^{3D} = (nf_k^{3D} + f_j^{2D})/(n+1)$, where n is the current number of fused observations for that instance. We also periodically prune instances below a size threshold defined by the median instance size and connect nodes that are spatially proximate. The resulting online scene graph $\mathcal{M}$ is built incrementally and matched against the BIM graph.

C. Semantic Monte Carlo Localization

We formulate S-MCL as a particle filter (PF) [16] with semantic-based measurement reweighting. At initialization, we sample M particles $\mathcal{X} = \{(\mathbf{x}_t^{[i]}, w_t^{[i]})\}_{i=1}^M$ over navigable BIM regions to approximate the global pose hypothesis. Each particle state is $\mathbf{x}_t^{[i]} = [x_t^{[i]}, y_t^{[i]}, \theta_t^{[i]}]^\top$, where $w_t^{[i]}$ denotes the particle weight. At the prediction step, given a control input $\mathbf{u}_t = (v_t, \omega_t)$ (linear and angular velocities), the particles are propagated according to a kinematic motion model $P(\mathbf{x}_t^{[i]}|\mathbf{x}_{t-1}^{[i]}, \mathbf{u}_t)$ perturbed with Gaussian noise $\epsilon_{\text{trans},t}^{[i]} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{trans}}^2\mathbf{I})$ and $\epsilon_{\text{rot},t}^{[i]} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{rot}}^2\mathbf{I})$.

To correct localization errors, we associate the current online observations with the BIM context under each particle hypothesis. First, each particle queries a precomputed offline visible-object set by ray casting within a sector defined by radius r and field of view (FOV), producing a candidate set of offline visible objects. Since offline BIM objects do not provide point-cloud features or visual embeddings, we use CLIP text embeddings extracted from category labels to compute cosine similarity with the online visual embeddings, while we introduce a spatial arrangement similarity to address the repetitive semantics issue. Specifically, both online observations and offline candidates are represented in a unified polar-coordinate view, where each object is described by the distance d and relative bearing angle α with respect to the current pose. As such, we apply a semantic-spatial coupled measurement model. Candidate pairs are first screened by semantic consistency using CLIP similarity, $\phi_{\text{sem}} = \cos(f_j^{3D}, C_k)$. For each retained pair, we define the spatial similarities of relative distance and angles as:

$$\phi_{\text{dist}} = \exp\left(-(|d_t^{\text{obs}} - d_t^{[i]}|)/\sigma_d\right), \tag{1}$$

$$\phi_{\text{ang}} = \exp\left(\kappa_\alpha(\cos(\alpha_t^{\text{obs}} - \alpha_t^{[i]}) - 1)\right), \tag{2}$$

The spatial score is then defined as $\phi_{\text{spa}} = \phi_{\text{dist}}\phi_{\text{ang}}$, and the final pairwise score ϕ_{tot} is obtained by combining semantic and spatial consistency together:

$$\phi_{\text{tot}} = \phi_{\text{sem}}\phi_{\text{spa}}. \tag{3}$$

This multiplicative approach penalizes inconsistency in any of the three terms more effectively than a weighted linear combination. When multiple objects are detected, we use the Hungarian algorithm [17] to enforce one-to-one optimal assignment between online observations and offline visible candidates in a polynomial time complexity, then average the resulting per-particle matching scores. The averaged score is converted into a likelihood through temperature scaling, and particle weights are updated:

$$w_t^{[i]} = w_{t-1}^{[i]} P(z_t|\mathbf{x}_t^{[i]}), \qquad w_t^{[i]} = \frac{w_t^{[i]}}{\sum_{k=1}^M w_t^{[k]}}. \tag{4}$$

TABLE I
QUANTITATIVE LOCALIZATION RESULTS OF *CoarseBIMLoc* ACROSS SIX TRAJECTORIES IN GAZEBO SIMULATION.

Traj	Trans. (m)	Orient. (°)	Room Acc. (%)
1	$0.60_{\pm 0.71}$	$12.0_{\pm 16.1}$	$96.0_{\pm 26.1}$
2	$0.71_{\pm 0.76}$	$15.3_{\pm 30.9}$	$90.0_{\pm 27.1}$
3	$0.75_{\pm 0.38}$	$14.9_{\pm 25.6}$	$92.0_{\pm 28.1}$
4	$0.32_{\pm 0.22}$	$4.3_{\pm 6.6}$	$94.0_{\pm 23.7}$
5	$0.58_{\pm 0.46}$	$18.8_{\pm 25.3}$	$90.0_{\pm 28.8}$
6	$0.62_{\pm 0.51}$	$14.5_{\pm 23.3}$	$94.0_{\pm 27.1}$
Avg.	$\mathbf{0.60}_{\pm 0.71}$	$\mathbf{12.0}_{\pm 16.1}$	$\mathbf{92.7}_{\pm 26.9}$

We trigger resampling when the effective sample size (ESS) falls below a threshold δ_{ess}, using cumulative-distribution sampling followed by weight reset. The final global pose is estimated as the weighted mean of particle states, $\hat{g}_t = \sum_{i=1}^M w_t^{[i]}\mathbf{x}_t^{[i]}$, where circular averaging is applied to yaw angles to avoid discontinuity. The corresponding room ID is obtained by querying the offline room partition at the estimated pose. This prediction-update-resampling cycle repeats until the particle set converges to a coarse global pose.

III. EXPERIMENTAL RESULTS

A. Implementation Details

Experimental setup. We evaluate CoarseBIMLoc in Gazebo simulation. We use a BIM model corresponding to one floor of a large real-world building, covering approximately $86 \times 86\,\text{m}^2$ in planar dimensions with a modeled height of $1.2\,\text{m}$, and containing $N = 7$ rooms and 43 common objects across 14 semantic categories. Room categories include bedroom, kitchen, meeting room, living room, study room, and hallway, while object categories include bed, chair, table, window, bookshelf, water closet, sink, and fire hydrant, and so on. The robot platform is a Clearpath Jackal equipped with an RGB-D sensor and wheel odometry. We conduct six trials in the BIM scene, each with a distinct initial pose and a trajectory length of $T = 50$, and repeat each experiment three times. Estimated global poses $\hat{g}_{1:T}$ are recorded for evaluation.

Hyperparameters. For online perception, we set both the box threshold δ_{box} and the text threshold δ_{text} to 0.25. For instance fusion, we set both δ_{sim} and δ_{IoU} to 0.7. For the particle filter, we set the number of particles to $M = 1000$, the FOV to $90°$, the ESS threshold to $\delta_{\text{ess}} = 0.5$, the translation noise to $\sigma_{\text{trans}} = 0.2$ m, and the rotation noise to $\sigma_{\text{rot}} = 10°$. We set spatial similarity coefficients to $\sigma_d = 1.2$ and $\kappa_\alpha = 5$.

Metrics. We calculate the average trajectory-wise pose error between ground-truth and estimated global poses to measure translation and orientation accuracy. The average error is defined as:

$$\text{Err} = \frac{1}{T}\sum_{t=1}^{T} \|g_t - \hat{g}_t\|_2, \tag{5}$$

where g_t denotes either the ground-truth xy position or yaw angle at timestep t, $\hat{g}_t$ denotes the corresponding estimate, and T is the trajectory length. We also compute room-level localization accuracy by comparing the predicted current room

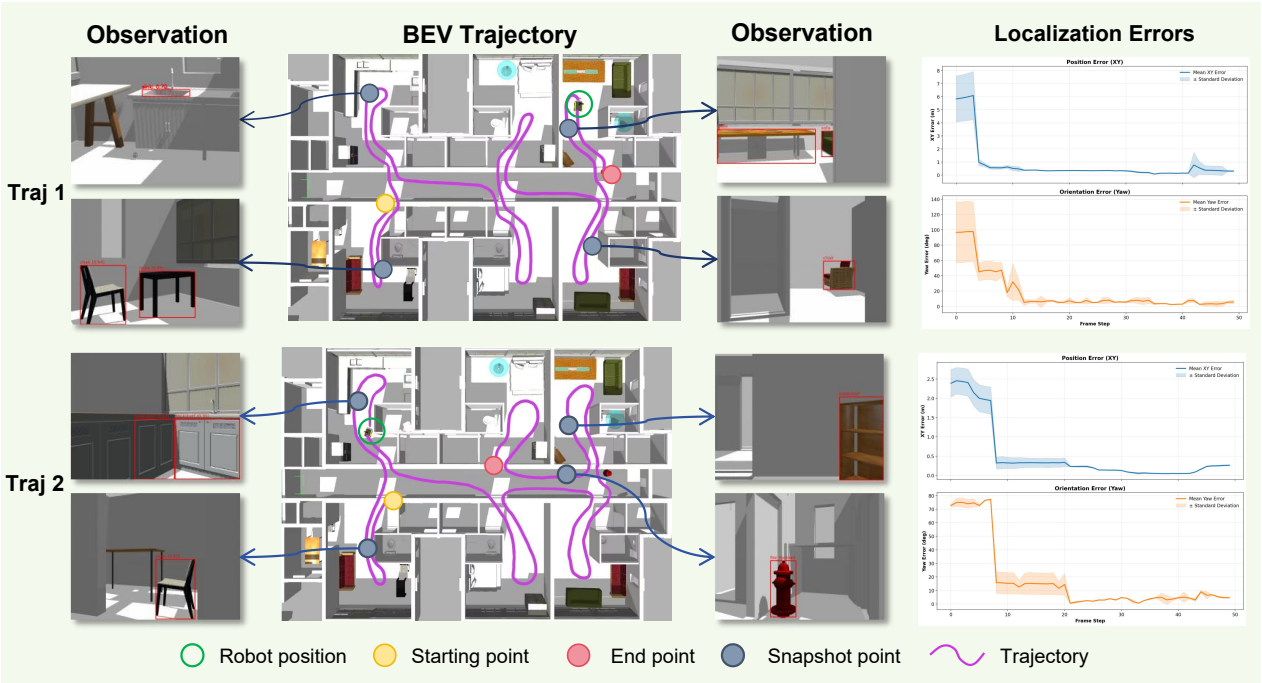

Fig. 2. Qualitative localization results of *CoarseBIMLoc* in Gazebo simulation. The figure depicts two representative trajectories, each repeated three times, illustrating the estimated global trajectory alignment against the ground truth and the visualization of particle convergence during relocalization. The performance demonstrates that iterative semantic and spatial matching effectively resolves pose ambiguity, avoids dependence on synthetic image style, and provides a robust coarse pose prior for fine-grained registration and navigation.

IDs with the ground-truth room IDs. The translation error (denoted as Trans. (m)), orientation error (denoted as Orient. (°)), and room-level localization accuracy (denoted as Room Acc.) are reported in Table I, validating that our method provides reliable pose priors in different settings.

B. Results and Analysis

Quantitative localization performance is summarized in Table I. Across six trajectories, each repeated three times, *CoarseBIMLoc* achieves an average translation error of $0.60 \pm 0.71\,\mathrm{m}$, an average orientation error of $12.0^\circ \pm 16.1^\circ$, and a room-level accuracy of $92.7\% \pm 26.9\%$. The translation error ranges from $0.32\,\mathrm{m}$ (Trajectory 4) to $0.75\,\mathrm{m}$ (Trajectory 3), and all trajectories remain below one meter. This indicates that the method consistently extracts stable coarse poses when navigating through BIM-derived environments.

A deeper data analysis shows that translation, heading, and room-level metrics capture different aspects of coarse localization quality. Trajectory 4 reports the lowest translation error ($0.32\,\mathrm{m}$) and orientation error (4.3°), while still maintaining high room accuracy (94.0%), suggesting that distinctive semantic-spatial layouts can support both metric and topological convergence. Trajectory 5 has the largest orientation error (18.8°), yet its translation error remains low at $0.58\,\mathrm{m}$, demonstrating that the semantic observation model reliably constrains the positional search space even when fine-grained yaw alignment is less stable.

The per-trajectory deviations in Table I (e.g., translation error $0.60 \pm 0.71\,\mathrm{m}$ in Trajectory 1 and orientation error $15.3^\circ \pm 30.9^\circ$ in Trajectory 2) further indicate that residual errors are concentrated in short ambiguous intervals instead of remaining uniformly high throughout the run. This pattern is consistent with a particle filter that starts with broad multi-modal hypotheses and then contracts once informative objects are observed. From a system perspective, the statistics support the intended role of *CoarseBIMLoc*: provide stable room-level and meter-level priors, then hand over fine orientation refinement to downstream geometric alignment.

Qualitative results in Fig. 2 are consistent with this interpretation. The particle set remains multi-hypothesis during early exploration and rapidly concentrates after observing distinctive room-specific object layouts. Despite the poor visual appearance of the BIM environment, semantic detectors naturally handle this disturbance and offer more distinctive semantic cues than image-level matching. The results also confirm the advantage of coupling semantics with spatial constraints: semantic similarity alone can be ambiguous for repeated categories (e.g., chairs or cabinets), while the polar-coordinate spatial consistency term suppresses implausible matches and improves reweighting stability. The combined quantitative and qualitative evidence validates the effectiveness of our method.

IV. Conclusion

We presented CoarseBIMLoc, a BIM-semantic localization framework based on open-set scene graph construction. By combining BIM-derived priors with real-time RGB-D semantic observations, the method delivers robust coarse localization in challenging indoor environments. The semantic Monte Carlo localization formulation reduces global pose ambiguity and provides an effective prior for downstream fine-grained geometric registration and navigation. Future work will integrate the algorithm with geometric registration using planar elements and pursue real-world deployment and evaluation in complex buildings with occlusions from decorative furniture.

REFERENCES

[1] U. Isikdag, S. Zlatanova, and J. Underwood, "A bim-oriented model for supporting indoor navigation requirements," *Computers, Environment and Urban Systems*, vol. 41, pp. 112–123, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0198971513000434

[2] S. Karimi, I. Iordanova, and D. St-Onge, "An ontology-based approach to data exchanges for robot navigation on construction sites," 2021. [Online]. Available: https://arxiv.org/abs/2104.10239

[3] L. Liu, B. Li, S. Zlatanova, and P. van Oosterom, "Indoor navigation supported by the industry foundation classes (ifc): A survey," *Automation in Construction*, vol. 121, 2021.

[4] Q. Li, R. Cao, J. Zhu, X. Hou, J. Liu, S. Jia, Q. Li, and G. Qiu, "Improving synthetic 3d model-aided indoor image localization via domain adaptation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 66–78, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621002677

[5] Y. Jia, J. Wang, W. Shou, M. R. Hosseini, and Y. Bai, "Graph neural networks for construction applications," *Automation in Construction*, vol. 154, p. 104984, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580523002443

[6] X. Zhao and C. C. Cheah, "Bim-based indoor mobile robot initialization for construction automation using object detection," *Automation in Construction*, vol. 146, p. 104647, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580522005179

[7] R. Zhai, J. Zou, V. J. Gan, X. Han, Y. Wang, and Y. Zhao, "Semantic enrichment of bim with indoorgml for quadruped robot navigation and automated 3d scanning," *Automation in Construction*, vol. 166, p. 105605, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580524003418

[8] Y. Zhang, L. Lu, X. Luo, and J. Pan, "Global bim-point cloud registration and association for construction progress monitoring," *Automation in Construction*, vol. 168, p. 105796, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580524005326

[9] C. Chen, L. Lu, L. Yang, Y. Zhang, Y. Chen, R. Jia, and J. Pan, "Signage-aware exploration in open world using venue maps," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3414–3421, 2025.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[12] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C. M. de Melo, J. B. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5021–5028.

[13] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," 2022.

[14] X. Li and J. Li, "Angle-optimized text embeddings," 2024. [Online]. Available: https://arxiv.org/abs/2309.12871

[15] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," 2024. [Online]. Available: https://arxiv.org/abs/2401.17270

[16] S. Thrun, D. Fox, W. Burgard, and F. Dellaert, "Robust monte carlo localization for mobile robots," *Artificial Intelligence*, vol. 128, no. 1-2, pp. 99–141, 2001.

[17] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.