Signage-Aware Exploration in Open World using Venue Maps

Chang Chen, Liang Lu†, Lei Yang†, Yinqiang Zhang, Yizhou Chen, Ruixing Jia, Jia Pan†

***Abstract*—Current exploration methods struggle to search for shops or restaurants in unknown open-world environments due to the lack of prior knowledge. Humans can leverage venue maps that offer valuable scene priors to aid exploration planning by correlating the signage in the scene with landmark names on the map. However, arbitrary shapes and styles of the texts on signage, along with multi-view inconsistencies, pose significant challenges for robots to recognize signage accurately. Additionally, discrepancies between real-world environments and venue maps hinder the integration of text-level information into the planners. This paper introduces a novel signage-aware exploration system to address these challenges, enabling robots to utilize venue maps effectively. We propose a signage understanding method that accurately detects and recognizes the texts on signage using a diffusion-based text instance retrieval method combined with a 2D-to-3D semantic fusion strategy. Furthermore, we design a venue map-guided exploration-exploitation planner that balances exploration in unknown regions using directional heuristics derived from venue maps and exploitation for perceiving the signage actively. Experiments in large-scale shopping malls demonstrate our method's superior signage recognition performance and search efficiency, surpassing state-of-the-art text spotting methods and traditional exploration approaches. Codes and videos are on our project website: sites.google.com/view/signage-aware-exploration.**

I. Introduction

Humans can efficiently navigate and explore a mall to search for shops or restaurants using a 2D venue map provided by the mall or Google Maps, even if the venue map is non-metric and only illustrates relative relationships between the different landmarks. In contrast, robots struggle to search for landmarks (here we refer to shops or restaurants) in unknown environments due to a lack of scene priors. However, this capability is fundamental for various applications, such as delivery, tour guidance, and inspection. Recently, significant efforts have been made to improve robots by mimicking human behaviors by correlating the visual observations with venue maps for global localization (e.g., Wang et al. [1] and SNAP [2]), or by discovering non-metric heuristics derived from maps for kilometer-scale navigation (e.g., ViKiNG [3] and S2MAT [4]). However, these approaches primarily focus on using geometric and semantic information from the venue maps but overlook the landmark names portrayed on the maps and the corresponding signage displaying the names in the scenes. Moreover, they often neglect online scene understanding and exploration. On the other hand, current exploration methods generally integrate geometric information [5], [6] or object-level scene semantics (e.g., Conceptgraphs [7] and HOV-SG [8]) but rarely incorporate text-level semantics, limiting their effective use of venue maps. Observing that humans exploit these *landmark names on the venue maps* by matching them to the *signage in the real-world scene*, we argue that a signage-aware exploration method is essential to improve a robot's ability to search for landmarks using 2D (non-metric) venue maps.

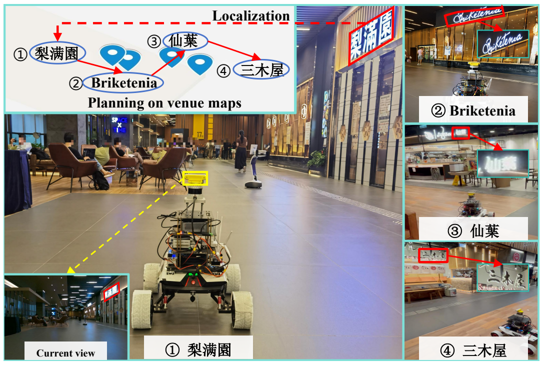


Fig. 1: We propose to leverage the *textual* information in a venue map to facilitate shop searching in unknown open-world environments. The robot localizes itself in the environment by recognizing and matching the texts on a sign to the venue map. Then the robot plans a direction to the next landmark 'Briketenia'.

However, implementing such a system poses the following key challenges. First, recognizing the landmark names on the signage in an open-world environment is inherently difficult. Existing works that leverage the textual cues as the landmarks for visual localization [9]–[11] often rely on optical character recognition (OCR) models trained on closed-set shapes and styles, which struggle to recognize signage effectively in open-world scenarios, where signage can exhibit diverse shapes, styles, and multi-view inconsistencies. Second, the accuracy of signage recognition also depends significantly on the robot's distance and orientation relative to the signage during movement, which requires an active perception policy to obtain clear observations. Third, using venue maps to guide exploration planning is challenging due to their inconsistent scales and distortions compared to real-world situations.

This paper presents a robotic exploration approach that leverages the signage in an unknown environment and the corresponding (non-metric) venue map to facilitate exploration and searching for landmarks. We propose a human-like retrieval process based on the appearance of signage to allow a closed-set detector to adapt to open-world situations without

†This work was partially supported by the Innovation and Technology Commission of the HKSAR government under the InnoHK initiative. (Corresponding authors: Liang Lu, Lei Yang, Jia Pan)

Chang Chen, Liang Lu, Lei Yang, Yizhou Chen, and Jia Pan are with Centre for Transformative Garment Production, Bldg 19W, Hong Kong Science Park, Hong Kong, China, and also with the Department of Computer Science, The University of Hong Kong, Pok Fu Lam, Hong Kong, China. (changchen@connect.hku.hk, llu92@hku.hk, l.yang@transgp.hk, yz.chen@transgp.hk, jpan@cs.hku.hk)

Yinqiang Zhang and Ruixing Jia are with the Department of Computer Science, The University of Hong Kong, Pok Fu Lam, Hong Kong, China. (zyq507@connect.hku.hk, ruixing@connect.hku.hk)

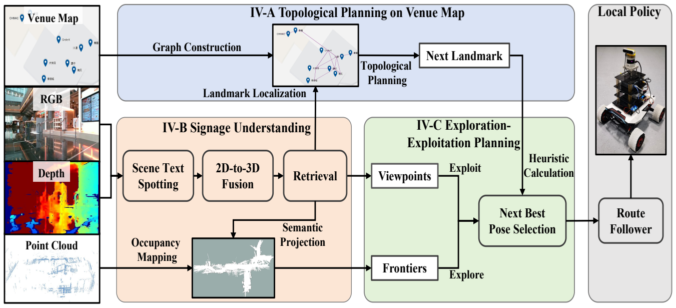


Fig. 2: Overall framework.

fine-tuning. It detects the texts on the signage and matches them with a set of pre-generated signage images based on the landmark names extracted from the given venue map. We also project multi-view 2D text regions into 3D space and fuse the corresponding features to enhance recognition performance. Furthermore, we design a venue map-guided exploration-exploitation planner that enables the robot to search for the signage corresponding to the landmarks using a directional heuristic. Once a sign candidate is detected, the system will exploit the known space to approach the sign and adjust the robot's view to faithfully recognize the sign. By balancing exploration and exploitation, we achieve both high signage coverage rates and search efficiency.

II. Methodology

We aim to design a signage-aware exploration method that leverages the venue map for searching for all the landmarks in unknown, large-scale, human-populated environments, such as shopping malls, thereby demonstrating our method's capability of navigating to any destination quickly. The environment contains N static landmarks (shops or restaurants), with corresponding signs displaying their landmark names$\mathcal{T}$. The venue map $\mathbf{M}$ (see Fig. 1) portrays all the landmarks. Our method first constructs a topological graph on a given venue map (Sec. II-A). Then, given the RGB-D image, the proposed signage understanding method recognizes the texts on the signage and correlates them with the text set of the venue map (Sec. II-B). Once localized on the venue map, the next landmark goal is inferred to guide the selection of frontiers. Our system balances exploration and exploitation to improve both signage coverage rates and search efficiency during the process (Sec. II-C). The overall framework is illustrated in Fig. 2.

A. Topological Planning on Venue Maps

We first pre-build a topological graph $\mathcal{G}$ based on the given venue map, whose nodes are the landmark names detected by an OCR model and whose edges connect two nodes. As such, we solve a travel salesman problem (TSP) to obtain a landmark route $g_{1:K}$. During online exploration, our method searches for the remaining landmarks sequentially on the topological graph to handle long-horizon planning [12].

B. Signage Understanding

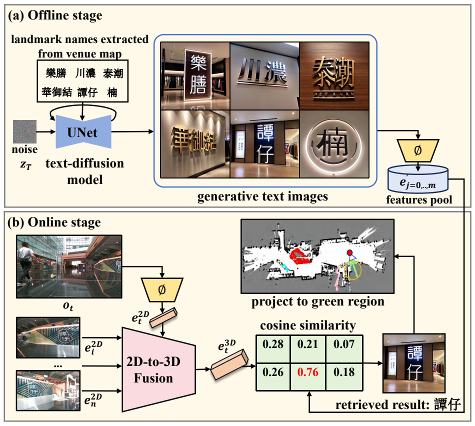


Fig. 3: The pipeline of signage understanding.

We online detect and recognize the signage during exploration, which enables the robot to globally localize itself on the venue maps. However, current OCR or STS models are limited by their closed-set training that cannot handle signage recognition with arbitrary shapes, styles, and multi-view inconsistency issues in the open world. To address the open-set recognition issue, we leverage the text set of the venue maps to perform visual similarity matching that enables the closed-set detectors to adapt to the open world without fine-tuning. Specifically, we employ a two-stage process for signage recognition. In the offline stage, we utilize an off-the-shelf multilingual text-diffusion model, AnyText [13], to convert all the landmark names $\mathcal{T}$ to the text-rendered signage images set, denoted as $\{o'\}$. Then, we adopt ESTextSpotter [14], denoted as ϕ, to detect text regions and only extract the text features e before the recognition heads from these offline-generated images without using the recognition results. This process creates a prior features pool $\mathcal{D}_e$. During online exploration, we utilize the same spotter ϕ to detect the signage from the current image observation o_t and extract its features e_t at timestep t. We then calculate the cosine similarity $s_\phi(o_t, o'_j)$ between the detected text features and those of the generative text images to find the most similar pair: $s_\phi(o_t, o'_j) = cos(\phi(o_t), \phi(o'_j))$. To mitigate the inconsistency issue of text recognition results from different views, we propose adopting a 2D-to-3D instance fusion strategy [15] to enhance recognition robustness using multiple observations. Using these fused features helps the robot to localize itself with the landmark names on the venue map. During exploration, we also construct a real-time signage map $\mathcal{M}$ for downstream querying and planning, which stores the recognized landmark names τ and the located regions. The signage understanding module is illustrated in Fig. 3.

C. Exploration-Exploitation Planning

During navigation to the next landmark g_t, the candidate next poses $c \in V$ are sampled among the set of unvisited frontiers $\mathcal{F}$ and unvisited viewpoints $\mathcal{V}$. The frontiers $f \in \mathcal{F}$ bias the exploration to unknown regions, while the viewpoints $v \in \mathcal{V}$ allow the robot to approach and face the signage for better recognition. In the initial stage, we induce the robot to explore its surroundings by selecting the frontiers that maximize the information gain $\sum G$ within the camera's FOV. After at least two landmarks are found, we align the online map with the venue map and estimate all the landmark poses p_{g_t} at the world coordinate using random sample consensus (RANSAC)

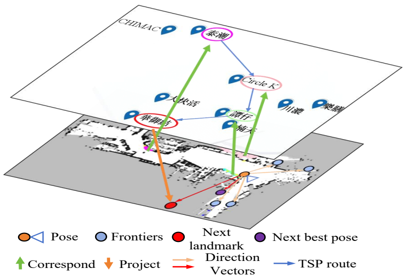


Fig. 4: Position estimation of the next landmark by calculating the coordinate transformation between the online map and the venue map for guiding the frontier selection.

[16]. Therefore, we perform the frontier-based informed search using the relative directions between landmarks on the venue maps. A directional heuristic $h(f_i)$ is calculated to favor selecting the frontiers closest to the direction towards the subgoal g_t. Therefore, the frontier utility $U_f(f_i, p_t)$ is computed as: $U_f(f_i, p_t) = \lambda\mu(p_t, f_i)h(f_i) - \eta d(p_t, f_i),\ \ f_i \in \mathcal{F}$.

On the other hand, the viewpoint utility $U_v(v_j, p_t)$ is defined to favor the exploitation in known spaces whenever a text instance with a certain confidence is detected: $U_v(v_j, p_t) = \beta s_v(v_j) - \eta d(p_t, v_j),\ \ v_j \in \mathcal{V}$. A factor β that determines to what extent of scores we should highlight the viewpoint candidates than frontiers is employed to handle the exploration-exploitation dilemma. Intuitively, the priority of exploiting signage originates from the fact that it can help localize the robot on the venue maps and facilitate the decision on the next best pose. Finally, we obtain the optimal next pose c_{t+1} among unvisited frontiers and viewpoints that maximizes the overall utility $U(c_j, p_t)$:

$$U(c_j, p_t) = I_f(c_j)U_f(c_j, p_t) + (1 - I_f(c_j))U_v(c_j, p_t). \quad (1)$$

where $c_j \in \mathcal{F} \cup \mathcal{V}$ and $I_f(c_j) = 1$ if $c_j \in \mathcal{F}$ else 0.

III. Experimental Results

A. Experimental Setup

We evaluate the proposed system in two large-scale shopping malls with 4 and 9 landmarks, respectively. We implement the system on a Scout-mini mobile robot platform shown in Fig. 2. We obtain the venue maps of the two scenarios via Google Maps, and we only showcase the landmarks of interest. We prompt AnyText [13] model to generate the text images by: *a sign of a store with* **[TEXT]** *written on it*, which means: 一个店铺标识，写着"**[TEXT]**" in Chinese, where **[TEXT]** is a place name extracted from the venue map using CnOCR.

B. Signage Recognition Performance

We evaluate our signage recognition method based on signage recognition recall rates on a signage image dataset collected in two scenarios, comparing it against four alternative methods: 1) using CLIP [17] to perform text-to-image similarity matching between landmark names and detected images, 2) using Chinese-CLIP [18], fine-tuned on Chinese dataset, 3) using recognition results of an OCR model, ESTextSpotter [14], with Levenshtein distance as the text-level measurement, and 4) font-based rendering retrieval (FontRR): leveraging the latent features extracted from ESTextSpotter and performing image-level retrieval with font-rendered text images [19]. Our method (DiffusionRR) utilizes a text-diffusion model for text-to-image generation. We collect the signage images of 18 shops and restaurants with 5 different views. The results are reported in Table I. It turns out that the signage recognition capabilities of both CLIP and Chinese-CLIP (ViT-B/16) are poor, achieving only 28.9% of recall@1. We attribute this to the inability of CLIP models to extract text features from signage images effectively. Moreover, the state-of-the-art STS method [14] still struggles with signage recognition, achieving only 52.2% recall rates due to the noisy recognition results. By converting texts to images using font (e.g., Arial) and calculating cosine similarity with signage images [19], the recall rates are improved by 10%. Furthermore, our diffusion-based approach yields an additional 15.6% improvement over the font-rending method.

Methods	Recall@1	Recall@2
CLIP [17]	28.9%	35.6%
Chinese-CLIP [18]	28.9%	45.6%
ESTextSpotter [14]	52.2%	52.2%
FontRR [19]	62.2%	73.3%
DiffusionRR (Ours)	**77.8%**	**87.8%**

TABLE I: Recall rates of signage recognition (↑)

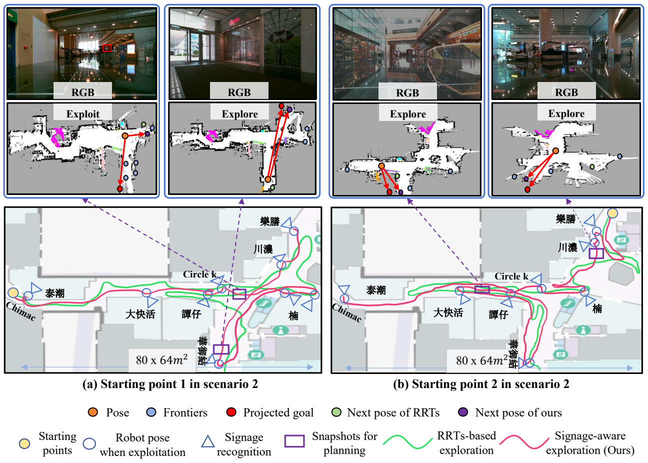
C. Signage Coverage Efficiency

This experiment is to evaluate the improvement in the efficiency of covering all the landmarks by using venue maps and exploitation method, comparing with RRTs-based exploration [5]. We also equip RRTs-based exploration with the proposed signage understanding module. We consider a landmark to be successfully covered if its corresponding signage is accurately recognized, thus here we use signage to represent landmarks. We choose two starting points in each scenario and conduct three trials at each point for two methods, respectively. We evaluate the signage coverage efficiency by both *signage coverage rates* and *exploration time per sign* with the standard variance. The former refers to the average number of recognized signs of all the signs in the scene, and the latter is calculated as $\frac{1}{n}\sum T(i)/S(i)$, where n is the number of trials, $T(i)$ and $S(i)$ are the total exploration time until no detected frontiers and signage coverage number of the i-th trial, respectively.

Scenario 1	Starting point 1		Starting point 2	
(4 landmarks)	Coverage rates ↑	Average time (s) ↓	Coverage rates ↑	Average time (s) ↓
[5]	1.50 ± 0.71 / 4	167.55 ± 25.51	1.67 ± 0.58 / 4	162.21 ± 70.67
Ours	**3.00** ± **0.00 / 4**	**67.89** ± **10.78**	**3.67** ± **0.58 / 4**	**74.94** ± **8.70**
Scenario 2	Starting point 1		Starting point 2	
(9 landmarks)	Coverage rates ↑	Average time (s) ↓	Coverage rates ↑	Average time (s) ↓
[5]	3.33 ± 0.58 / 9	186.67 ± 67.79	4.00 ± 1.00 / 9	171.55 ± 47.73
Ours	**7.00** ± **1.00 / 9**	**93.40** ± **14.23**	**6.67** ± **0.58 / 9**	**120.58** ± **30.22**

TABLE II: Coverage rates and exploration time per sign.

The experiment results for comparing our entire system with the baseline are reported in Table II, and the qualitative examples of the trajectories are illustrated in Fig. 5. We see that the RRTs-based method covers around 1 and 1.67 of 4 signs at two starting points in scenario 1, respectively. For the larger scenario 2, it can only find 3 and 4 of 9 signs. This



(a) Starting point 1 in scenario 2

(b) Starting point 2 in scenario 2

Fig. 5: Qualitative examples of the exploration trajectories in four scenario settings. Our method produces more efficient and reasonable exploration paths by using venue maps.

is because the RRTs-based method can only recognize the signs when the robot faces them with proper orientations while missing recognizing some signs. Our method successfully covers more signs (3 and 3.67 of 4 signs in scenario 1 and 7 and 6.67 of 9 signs in scenario 2) thanks to the exploitation behavior, enabling the robot to approach the signs and adjust the orientation for better recognition. On the other hand, in Fig. 5, the RRTs-based method shown in green trajectories blindly explores unknown regions with many redundant steps, yielding inefficient paths and spending about 168s and 162s in scenario 1 and 187s and 172s in scenario 2, respectively, for searching for each sign. By using venue maps, our method in red trajectories localizes the robots on the venue maps after recognizing the sign, thereby knowing the approximate locations of the next signs by calculating the directional heuristic. Therefore, our method navigates to the signs quickly within the narrowed regions, which results in a reduction of nearly 1x the exploration time per sign compared to the baseline and produces more efficient exploration paths to cover the signs (see Fig. 6).

D. Impact of Balance Weight

An important design choice is the weighting parameter β, which balances venue map-guided exploration and exploitation. To investigate the impact of venue maps and exploitation, we evaluate β with values $\{3, 9, 15\}$. A higher weight assigns greater importance to exploitation. When the weight is too low ($\beta = 3$), the planner is approximately degraded to the RRT-based exploration with only venue maps. While inclining to frontier exploration, it often fails to recognize certain signs due to poor observations. On the other hand, as the weight is too high ($\beta = 15$), the planner is approximately degraded to the RRT-based exploration with only exploitation that it prioritizes selecting viewpoints for perceiving the potential signage. A moderate balance of exploration and exploitation ($\beta = 9$ by default) can achieve both higher signage coverage rates and search efficiency.

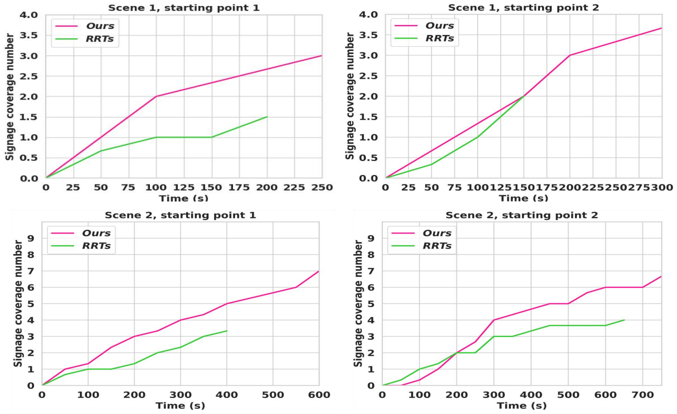


Fig. 6: The exploration progresses in four scenarios. The curves stop when the last new sign is covered, after which the exploration may continue but no more sign is covered.

IV. Discussion

1) Computational overhead. Our system relies on lightweight algorithms optimized for real-time applications. The text detection and retrieval processes run asynchronously, consuming approximately 5GB of GPU memory and achieving a speed of ∼2fps on our edge device (157 TOPS). This ensures high performance in signage understanding with minimal delay. **2) Scalability.** Our approach may be less effective in large scenes with sparse signage since the lack of signs can hinder localization. While our approach emphasizes the effectiveness of leveraging textual information in the scene, one can always integrate our approach into conventional ones to leverage the spatial structures for global localization. In scenes with dense signage, while false positives may increase, our system includes a RANSAC step to eliminate spatially distant matches. Additionally, dense signage can benefit text-absent issues, aiding in initial localization and exploration. **3) Multilingualism.** In our scenes, some signs display texts in multiple languages (e.g., Japanese and Korean), which may interfere with accurately recognizing target texts in Chinese and English. We find our method can handle this through appearance-based similarity matching, even though our text detector has not been trained on these additional languages. Fine-tuning the text detectors on multilingual datasets can further enhance multilingual recognition, as AnyText [13] supports rendering multilingual text on generative images.

V. Conclusion

We present the first signage-aware exploration method that leverages signage in the scenes and the 2D non-metric venue maps to incorporate text-level information for searching for landmarks in unknown open-world environments. To overcome the challenge of signage recognition, we proposed a diffusion-based text instance retrieval method that detects and recognizes signage with arbitrary shapes and styles in the scene effectively. A 2D-to-3D semantic fusion strategy is employed to enhance recognition performance. Furthermore, we design a venue map-guided exploration-exploitation planner to achieve both high signage coverage rates and search efficiency. Real-world experiments demonstrate that our method is more efficient and robust compared to the baselines.

References

[1] S. Wang, S. Fidler, and R. Urtasun, "Lost shopping! monocular localization in large indoor spaces," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2695–2703.

[2] P.-E. Sarlin, E. Trulls, M. Pollefeys, J. Hosang, and S. Lynen, "Snap: Self-supervised neural maps for visual positioning and semantic understanding," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] D. Shah and S. Levine, "Viking: Vision-based kilometer-scale navigation with geographic hints," *arXiv preprint arXiv:2202.11271*, 2022.

[4] T. Fan *et al.*, "S^{2}mat: Simultaneous and self-reinforced mapping and tracking in dynamic urban scenariosorcing framework for simultaneous mapping and tracking in unbounded urban environments," 2023.

[5] H. Umari and S. Mukhopadhyay, "Autonomous robotic exploration based on multiple rapidly-exploring randomized trees," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017, pp. 1396–1402.

[6] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon" next-best-view" planner for 3d exploration," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1462–1468.

[7] Q. Gu *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," 2023.

[8] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[9] N. Radwan, G. D. Tipaldi, L. Spinello, and W. Burgard, "Do you see the bakery? leveraging geo-referenced texts for global localization in public maps," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 4837–4842.

[10] N. Zimmerman, L. Wiesmann, T. Guadagnino, T. Läbe, J. Behley, and C. Stachniss, "Robust onboard localization in changing environments exploiting text spotting," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 917–924.

[11] B. Li, D. Zou, Y. Huang, X. Niu, L. Pei, and W. Yu, "Textslam: Visual slam with semantic planar text features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[12] C. Chen, Y. Liu, Y. Zhuang, S. Mao, K. Xue, and S. Zhou, "Scale: Self-correcting visual navigation for mobile robots via anti-novelty estimation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 16 360–16 366.

[13] Y. Tuo *et al.*, "Anytext: Multilingual visual text generation and editing," *arXiv preprint arXiv:2311.03054*, 2023.

[14] M. Huang *et al.*, "Estextspotter: Towards better scene text spotting with explicit synergy in transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 495–19 505.

[15] S. Lu, H. Chang, E. P. Jing, A. Boularias, and K. Bekris, "Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data," in *7th Annual Conference on Robot Learning*, 2023.

[16] D. Barath, J. Matas, and J. Noskova, "Magsac: marginalizing sample consensus," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 197–10 205.

[17] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[18] A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, and C. Zhou, "Chinese clip: Contrastive vision-language pretraining in chinese," *arXiv preprint arXiv:2211.01335*, 2022.

[19] L. Wen *et al.*, "Visual matching is enough for scene text retrieval," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 447–455.