

Language-enhanced RNR-Map: Querying Renderable Neural Radiance Field maps with natural language

Francesco Taioli^{1,*}, Federico Cunico^{1,*}, Federico Girella^{2,*},
Riccardo Bologna^{2,*}, Alessandro Farinelli², Marco Cristani¹

¹University of Verona, Dept. of Engineering for Innovation Medicine

²University of Verona, Dept. of Computer Science

name.surname@univr.it, name.surname@studenti.univr.it[†]

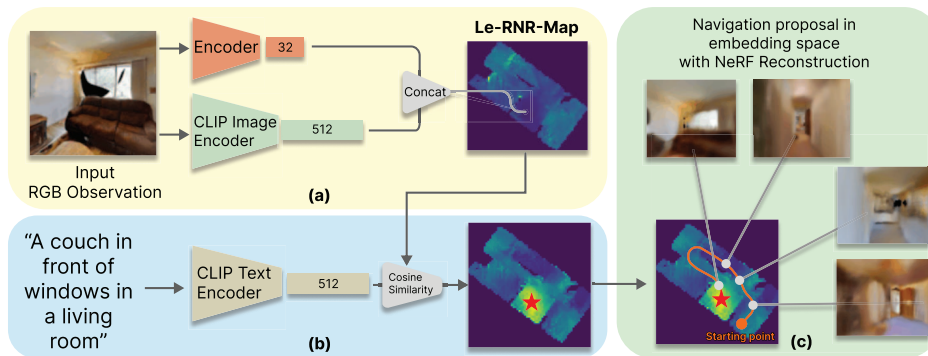


Figure 1: Overview of Le-RNR-Map. (a) shows Le-RNR-Map map construction in which we embed visual and visual-language-aligned features, (b) shows the query search, at inference time, with natural language, and (c) shows the reconstruction of the images on the path, from starting position (orange circle) to the goal (red star) using NeRF.

Abstract

We present *Le-RNR-Map*, a *Language-enhanced Renderable Neural Radiance map for Visual Navigation with natural language query prompts*. The recently proposed *RNR-Map* employs a grid structure comprising latent codes positioned at each pixel. These latent codes, which are derived from image observation, enable: *i) image rendering given a camera pose, since they are converted to Neural Radiance Field; ii) image navigation and localization with astonishing accuracy*. On top of this, we enhance *RNR-Map* with *CLIP-based embedding latent codes*, allowing natural language search without additional label data. We evaluate the effectiveness of this map in single and multi-object searches. We also investigate its compatibility with a Large Language Model as an “affordance query resolver”. Code and videos are available at the link <https://intelligolabs.github.io/Le-RNR-Map/>.

1. Introduction

Embodied AI is receiving a lot of attention in recent years, with interesting yet very challenging tasks such as

Embodied Question Answering [27], Image-Goal Navigation [13], Visual-Language Navigation [14] and zero-shot Object-Goal navigation [18, 8]. Currently, some areas of research are focusing on creating explicit map representations that could improve the performances on those tasks, such as semantic segmentation maps [2, 19], occupancy maps and top-down semantic map prediction [9]. More recently, some works attempted to embed latent vectors in the explicit map to perform navigation [11]. However, very few tried to combine Visual-Language-aligned and NeRF [20] latent codes into the map itself to solve different tasks simultaneously.

For this reason, in this preliminary work, we expand *RNR-Map* [15] by enhancing it with visual-language-aligned features. The result is a *Language enhanced, Renderable Neural Radiance Map (Le-RNR-Map)* for visual navigation that is *visually descriptive*, thanks to the Neural Radiance Field embeddings, *generalizable*, since it uses off-the-shelf models requiring no training, and queryable with *both text and images* thanks to the addition of visual-language-aligned features. To the best of our knowledge, this is the first work in the literature to create a map representation that allows to solve three distinct tasks *at the same time*: *i) Localization of objects given a query image, by using the RNR-Map [15] embeddings; ii) Open Vo-*

^{*}The authors contributed equally to this paper

cabulary localization of objects through natural language, using Visual-Language-aligned embeddings; *iii*) rendering of the path to the target, highlighting the searched object with Visual-Language-aligned features, without the need to physically move the agent.

2. Related Literature

2.1. Maps For Navigation

There is an extensive research area that focuses on how to build maps to aid navigation in indoor environments [1, 4, 19, 11, 15, 26]. An occupancy grid map is a $m \in \mathbb{R}^{M \times M \times C}$ matrix, where $M \times M$ is the spatial size and C is the number of channels storing information about a corresponding region. Recently, the authors of RNR-Map [15] introduced a novel type of map, in which the latent codes are embedded from visual observations and can be converted to a neural radiance field, which enables image rendering given a camera pose, thus being visually descriptive. Moreover, the author showed that this novel type of map can be useful for visual localization and navigation.

In AutoNeRF [19], the authors introduce a method to collect data required to train NeRFs using autonomous embodied agents, and use the experience to build an implicit map representation of the environment. Moreover, they augment the NeRF rendering procedure with a segmentation head over S predefined classes. In VLMaps [11], the authors ground language information to visual observation fusing pre-trained visual language features [16] into a 2D spatial representation. A similar approach is presented in [3]. In contrast to [11], Le-RNR-Map also allows us to perform Image-Goal Navigation, and NeRF rendering given a camera pose. Additionally, Le-RNR-Map can be built faster, in around 60 seconds, for ~ 1000 RGB-D images (128×128).

2.2. Nerf

Neural Radiance Fields, introduced in [20], address the problem of view synthesis, that is generating scene views from unseen novel viewpoints. NeRF scenes are modelled by a multilayer perceptron network which outputs the radiance emitted by a 3D point, given a spatial location (x, y, z) and a viewing direction (θ, ϕ) . The original NeRF formulation [20] can only represent small scenes, and does not generalize to new scenes/objects. To solve these limitations, other works take the challenge of learning a distribution over complex scenes. Generative Scene Networks (GSNs) [6] can be used to learn a rich scene prior in order to generate new scenes or fill the given one, decomposing the scene in local radiance fields that can be rendered from a moving camera. Moreover, they can be used to render images from latent codes, which in our case are stored in Le-RNR-Map, like the original RNR-Map formulation

[15]. Some recent works have extended the principle goal of NeRF with some other tasks. LERF [12] proposed a method for grounding CLIP representations in a dense, multi-scale 3D field, which can render dense relevancy maps given textual queries. However, LERF is still limited to small scenes and requires 45 minutes for a capture.

2.3. Language and Vision

Mapping text and images is the problem of estimating a function that maps images to the desired text (e.g. captioning [17]) and vice versa (e.g image generation with text-conditioning [23, 24]). Introduced in [22], CLIP is a model composed of an image-encoder and text-encoder trained to map embeddings from images and their description close to each other in the feature space, with a contrastive loss that enforces non-related pairs to be mapped further from each other. The authors showed with extensive experiments that CLIP achieves competitive zero-shot performance and thus can be used as a foundational model in a variety of task, such as scene segmentation [21, 16], Open-Vocabulary object detection [10] and Image-Generation [23]. Moreover, [7] introduces MaskCLIP, a framework for obtaining scene segmentation with indirect supervision from language.

3. Method

3.1. Map creation

We create a Le-RNR-Map by first extracting visual and Visual-Language-aligned features from RGB frames, while a subsequent feature registration process projects them in the map (Fig. 1a).

Visual embeddings. Inspired by [15] we consider a robot agent exploring the scene with a random walk, using RGB-D data and its on-board sensors, *i.e.* odometry information, to build Le-RNR-Map. An encoder-decoder architecture performs the creation of the RNR-Map. To allow an effective encoding of the 3D environment, the authors of [15] perform training of the encoder-decoder as follows: *i*) the encoder takes in input the RGB-D image and extracts the pixel features; *ii*) the decoder uses pixel features, along with the current pose (*i.e.* position of the agent), to sample latent information along each camera ray corresponding to each pixel, and tries to render the corresponding images. The latent codes extracted from the encoder, then, represent the pixel-level visual information from the current view. In our experiments, the encoding features are $F_{rnr} \in \mathbb{R}^{32}$. These features allow image rendering using Neural Radiance Field, and image localization in the map. For a more in-depth description, we refer the reader to [15].

Natural language. To include language features, we use a pre-trained CLIP [22] image-encoder to get $F_{clip} \in \mathbb{R}^{512}$ from the current RGB-D frame.

Le-RNR-Map. The final embedding space for the

current observation RGB-D frame is then composed as $F_{le-rnr} = F_{clip} \oplus F_{rnr}$. $F_{le-rnr} \in \mathbb{R}^{544}$ with the first 32 channels generated from RNR-Map and the 512 remaining from CLIP. Both the features from RNR and CLIP are then projected to the 2D map using the depth information, as in [15]. This allows us to keep the exact performances of RNR-Map for the Image-Goal navigation task, with the addition of being able to query the navigation through natural language thanks to the CLIP features.

3.2. Language-vision object search

The vision-language object search is performed by providing a natural language query (Fig. 1b). This query should indicate the objects required to find (either big furniture or small objects) that the navigation module has to handle as goal objects. Once the query is provided, the text embeddings are extracted with a pre-trained CLIP [22] text-encoder, and the cosine similarity with each cell of the Le-RNR-Map is computed. We select the location of maximal similarity as the goal location. The 3D end-goal predicted location, given the (x, y) indices expressed in Le-RNR-Map coordinate system, is obtained through inverse projection as in [15]. Exact procedure is presented in the supplementary material. To ensure the correct maximal similarity is found, prompt engineering is performed with negative prompting following [12, 22]. Together with the query prompt, this empirically shows a more fine-grained similarity on the maps and gives better localization as shown in Fig. 3. An extension of GSN [6] is then used to synthesize novel views and render a possible pathway that leads from any point starting point of the map to the required goal. Once the navigation reaches the proximity of the goal (*i.e.* the location of maximal similarity in the Le-RNR-Map), we estimate the camera orientation towards the goal by rotating the camera by 360° , computing the CLIP features for each degree and choosing the one with maximal cosine similarity with the target query. Finally, the visual saliency of the goal (Fig. 2c) is extracted from the RGB reconstructed by GSN [6] using pixel-wise CLIP features [7]. When multiple target objects are requested, the navigation is performed sequentially for each object following the same procedure, using as starting location the previous target location.

4. Experiments

To highlight the benefits that Le-RNR-Map brings to the Object-Goal Navigation task, we designed a set of targeted experiments. In Sec. 4.1 we evaluate the ability of Le-RNR-Map to correctly locate items using only Natural Language prompts. In Sec. 4.2 we show that the Renderable Neural Radiance map allows the user to properly select the correct item of interest in case of ambiguity. Finally, in Sec. 4.3, we explore the possible collaboration between a

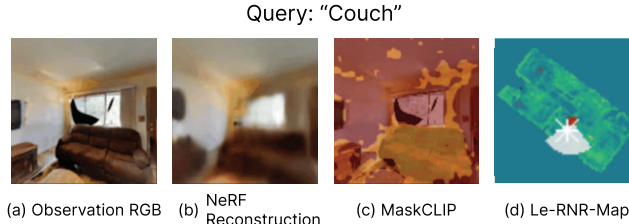


Figure 2: (a) Observation from Habitat-sim [25]. (b) Reconstruction using latent code from Le-RNR-Map using Neural Radiance Field (c) Feature visualization of the text query using MaskCLIP [7]. (d) Top-down view of Le-RNR-Map.

Large Language Model and Le-RNR-Map to find items and locations based on contextual prompts, called *Affordance queries* (*e.g.* the query “Find me a drink to wake me up” results in the agent looking for a cup of coffee). All of our experiments were conducted inside the Habitat-Sim [25] using the Gibson [28] dataset.

4.1. Searching items by Language prompts

The goal of Le-RNR-Map is to provide the user with a Natural Language interface with the agent. Such an interface would allow the user to prompt the agent with low effort, even in a constrained environment where traditional interactions may be unavailable (*e.g.* the user is holding something and can’t physically interact with the agent) or improbable (*e.g.* asking an agent to look for a particular object by showing it a picture of the object itself). With this goal in mind, we test the ability of Le-RNR-Map to locate different common items and/or locations in the environment, using only the CLIP features embedded in the navigation map.

In Tab. 1 we report the Success Rate and Distance To Success (DTS), as defined in [2], on some scenes of the validation split of the Gibson tiny dataset [28]. The dataset provides a textual label for the target object and its location in the scene as ground truth. We use the label as text prompt and compute the metrics comparing our predicted location with the ground truth. Additionally, as explained in Sec. 3.2, we define a series of unwanted objects or general/background elements (*e.g.* *stuff*, *wall*, *floor*) as negative prompts. These prompts may be specific for each scene. Together with the target prompt, we compare their similarity results with the map embeddings (as seen in [22]) resulting in similarity maps with more consistent areas of interest as shown in Fig. 3. In general, Le-RNR-Map enables us to obtain a decent success rate and DTS without additional training required. Moreover, we investigate the low Success Rate for the *Darden* scene. We found that the observations, given to the agent during the map creation, contain several artefacts, such as mirror reflections, missing walls and mesh holes, leading to incorrect embeddings into the map. Further study will analyze this problem.

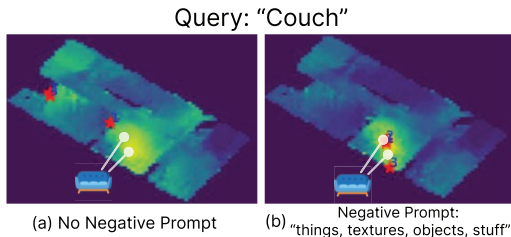


Figure 3: Similarity heatmap between the prompt `Couch` and Le-RNR-Map. (a) is without negative prompts. (b) shows that negative prompts result in cleaner maps with more concentrated similarity zones. The stars indicate the maximum similarity locations.

Table 1: Results on the `val` split of Gibson tiny dataset. Note that our setup is *known* since we generate the map beforehand. Each scene has a different negative prompt.

Scene name	Negative Prompts	Success \uparrow	DTS (m) \downarrow
<i>Corozal</i>	\times	0.69	1.19
	\checkmark	0.69	0.65
<i>Darden</i>	\times	0.37	4.12
	\checkmark	0.59	2.59
<i>Markleeville</i>	\times	0.81	1.38
	\checkmark	0.83	1.50
<i>Wiconisco</i>	\times	0.76	0.50
	\checkmark	0.76	0.50
Average	\times	0.66	1.79
	\checkmark	0.72	1.31

4.2. Solving prompt ambiguities

One of the advantages of having a Renderable Neural Map is the possibility for the agent to *explore the scene without actually moving in the real world*. This is particularly useful in scenarios where the user prompt may be ambiguous and refer to multiple objects or locations in the scene (e.g. `window` may refer to different windows). When this happens, the RNR-Map can be used to provide the user with visual previews of the paths it would take to get to all the possible solutions. In this section, we explore this case and provide a solution using Le-RNR-Map.

First, as in Sec. 4.1, we compute similarities between the CLIP features extracted from the prompt and the ones embedded in Le-RNR-Map by also using negative prompts. After finding the maximum similarity location, we suppress all the similarities in the adjacent cells of the map. We then look for the new maximum similarity above a threshold $th = 0.6$. For each target found, we render the path that the agent would follow to reach the target using a shortest path algorithm. The process ends when there is no longer a similarity score greater than th . We consider it a success when this process finds the target item in at least one

of the N predicted paths, simulating the user “selecting” the desired item. While this evaluation is heavily reliant on hyper-parameters, such as the number N of predicted paths allowed and the value T of the threshold, we still believe it to be an interesting use case, and propose our work as a first informal approach to this problem. For video examples, we refer the user to the supplementary material.

4.3. Affordance search

We argue that Natural Language alone is not enough to achieve natural interaction with the user and the agent, especially if we restrict the user to a limited set of words (classes) or a rigid sentence structure. The idea comes from the following observation: what if we want to search for some specific location of an indoor environment, but we are unable to express the query in a direct way? As an example, consider the scenario where we want to search for a location “*that can be relaxing after a long day at work*”. We define this use case as “affordance search”, and propose to take advantage of the current state-of-the-art Large language model to translate the query and output a set of possible target descriptions, using the available GPT-3.5 chat-completions API. After retrieving the descriptions, we sequentially search for each target with the procedure presented in Sec. 3.2. Visualizations and details about prompts are available in the supplementary material.

5. Conclusions & Future works

In this preliminary study, we have enhanced the novel RNR-Map [15] to allow natural language search using an off-the-shelf model. We qualitatively show the result using single and multi-object search, generating videos of the shortest path to reach the object using the neural Radiance Field from latent code inside RNR-Map. Moreover, we show how LLMs can be used as an “affordance query resolver”. Several interesting future works could follow this preliminary study. We plan to improve the rendering quality of the reconstructed observation by adding a Language-driven grounding head to the NeRF procedure, similar to [12] but in an open-scene case. Research could also focus on training an end-to-end agent to solve the zero-shot object goal navigation, and studying if the online generation of Le-RNR-Map serves as an auxiliary task to improve the generalization capabilities of the agent. Furthermore, we plan to evaluate the impact of different negative prompts in the standard zero-shot object goal navigation benchmark. Also, we plan to study how to deal with dynamic environments, thus providing a way to update the map, both in NeRF and language embedding space. Finally, we are outlining a real-world implementation in an industrial environment, leveraging Human-Robot-Interaction methods [5] for better integration of human-guided robot navigation and enhanced task efficiency.

References

- [1] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020.
- [2] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020.
- [3] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *arXiv preprint arXiv:2209.09874*, 2022.
- [4] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.
- [5] Federico Cunico, Marco Emporio, Federico Girella, Andrea Giachetti, Andrea Avogaro, and Marco Cristani. Oo-dmvm: A deep multi-view multi-task classification framework for real-time 3d hand gesture classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2744–2753, 2023.
- [6] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE International Conference on Computer Vision*, page 14284 – 14293, 2021.
- [7] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.
- [8] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023.
- [9] Georgakis Georgios, Schmeckpeper Karl, Wanchoo Karan, Dan Soham, Miltsakaki Eleni, Roth Dan, and Daniilidis Kostas. Cross-modal map learning for vision and language navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.
- [11] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.
- [12] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lurf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023.
- [13] Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-specific image goal navigation: Training embodied agents to find object instances. *arXiv preprint arXiv:2211.15876*, 2022.
- [14] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 104–120. Springer, 2020.
- [15] Obin Kwon, Jeongho Park, and Songhwa Oh. Renderable neural radiance map for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9108, 2023.
- [16] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.
- [17] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23359–23368, 2023.
- [18] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [19] Pierre Marza, Laetitia Matignon, Olivier Simonin, Dhruv Batra, Christian Wolf, and Devendra Singh Chaplot. Autonerf: Training implicit scene representations with autonomous agents. *arXiv preprint arXiv:2304.11241*, 2023.
- [20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [21] Songyou Peng, Kyle Genova, Chiyu “Max” Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
- [23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans,

- et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- [25] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [26] Francesco Taioli, Francesco Giuliari, Yiming Wang, Riccardo Berra, Alberto Castellini, Alessio Del Bue, Alessandro Farinelli, Marco Cristani, and Francesco Setti. Unsupervised active visual search with monte carlo planning under uncertain detections. *arXiv preprint arXiv:2303.03155*, 2023.
- [27] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [28] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.