Proceedings of 2023 IEEE
International Conference on Mechatronics and Automation
August 6 - 9, Harbin, Heilongjiang, China

A Survey of Object Goal Navigation: Datasets, Metrics and Methods

Dewei Wang, Jiaming Chen, Jiyu Cheng*
The School of Control Science and Engineering
Shandong University
Jinan, China
** Corresponding Author*
dweik@mail.sdu.edu.cn, ppjmchen@gmail.com, jycheng@sdu.edu.cn

2023 IEEE International Conference on Mechatronics and Automation (ICMA) | 979-8-3503-2084-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMA57826.2023.10215817

***Abstract*—Object Goal Navigation (ObjectNav) aims at directing an agent to a specified target object within an unseen scene. This task integrates advanced techniques, including visual perception, semantic prior learning, and visual navigation, which makes it a challenging and emerging task within the Embodied AI field. In this survey paper, we offer a comprehensive review of widely-used datasets and corresponding evaluation metrics for ObjectNav. Additionally, we provide an in-depth review and analysis of recently developed methods. Finally, we present discussions on potential future research directions.**

***Index Terms*—Object Goal Navigation, Embodied AI, Deep Learning, Reinforcement Learning**

I. Introduction

Embodied AI [1] has recently experienced substantial advancements [2], establishing itself as a thriving research field that emphasizes the integration of AI techniques, such as computer vision and natural language processing, with physical entities. Embodied navigation, a fundamental topic within the Embodied AI domain, focuses on guiding an agent towards a target goal in an unfamiliar environment. In accordance with the classification of navigation goals, Anderson [3] categorizes Embodied Navigation tasks based on their objectives as follows: Point Goal Navigation [4] (a specific coordinate, e.g., *Go 5m north, 3m west relative to start*), Object Goal Navigation [5] (a particular object instance, e.g., *cabinet*), Area Goal Navigation (a specific region, e.g., *kitchen*), and Language Navigation [6] (text describing the navigation routine, e.g., *walk past the piano through an archway directly in front*). Among these tasks, we concentrate on Object Goal Navigation (ObjectNav) in this survey paper, as it presents unique challenges and opportunities for research.

As illustrated in Fig. 1, the agent is initialized randomly within an unseen environment and is equipped with sensors, including a camera, GPS, and compass, as well as a specific target object category. The objective of this task is to navigate to the target object by executing a sequence of actions, such as ”*FORWARD*”, ”*TURN LEFT*” and ”*TURN RIGHT*”, while exploring the environment. The success of this task is determined by the agent taking a ”*STOP*” action and being in close enough to any instance of the target object category. ObjectNav requires agents to exhibit not only robust navigation skills, but also the ability to recognize and locate specific objects within dynamic and complex environments. It serves as an excellent testbed for developing and evaluating innovative techniques in Embodied AI, ultimately contributing to the advancement of the field.

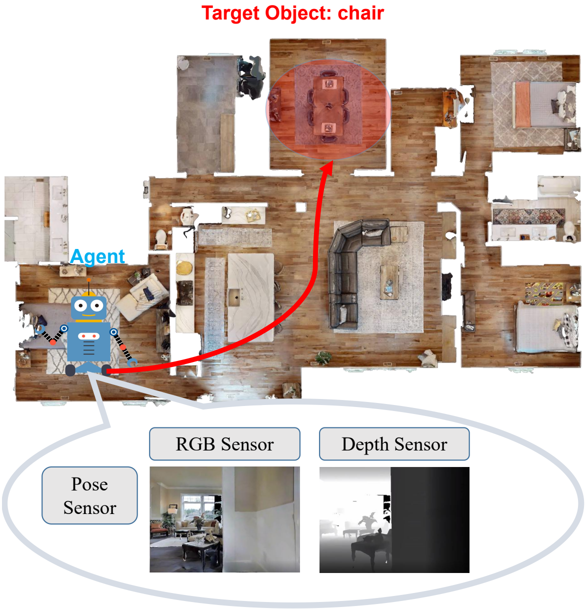

Fig. 1. In Object Goal Navigation, the agent is initialized at a random location within an unfamiliar environment and is required to navigate to a specific target instance. Throughout the navigation, observations from various sensors, including RGBD image and pose (from compass and GPS), are provided at each step.

In this survey paper, we conduct an in-depth analysis of existing datasets, online challenges, and recently proposed methods. We gathered information from published papers, datasets [7]–[11], as well as completed challenges [5], [11], [12]. The remainder of this paper is organized as follows: Section II discusses the datasets, while Section III addresses the metrics. Section IV provides an overview and analysis of the recently proposed methods and their results in the challenges. Finally, Section V offers a brief discussion on potential future research directions on ObjectNav.

II. Datasets

In this section, we present a review of widely-used datasets on ObjectNav, which can be further divided into public datasets and online challenge benchmarks.

A. Public Datasets

We review four commonly-used public datasets [7]–[11] for ObjectNav as follows:

1) Gibson: Gibson [7] is the largest building-level reconstruction dataset, which is based on virtualizing real spaces, rather than artificially designed ones. The Gibson databaseincludes over 1400 floor spaces from 572 full buildings. [13] devise a semi-automatic framework that employs existing detection methods to follow the Scene Graph paradigm in 3D in order to construct a graph that spans the entire building and includes semantics on objects, rooms and cameras, as well as the relationships among these entities providing object segmentations and labels on 35 homes. This generate a dataset that provides a 25/5/5 split for training, validation and testing, which is usually used for ObjectGoal Navigation task.

2) Matterport3D: Matterport3D (MP3D) [8] is a large-scale 3D scene dataset, consisting of 90 building-scale scenes with 10,800 panoramic views from 194,400 RGB-D images. There are a total of 50,811 object instance annotations in the 3D segmentations and 1,659 unique text labels, which are then standardized to a set of 40 object categories after cleaning. Each scanned building, on average, has 2.61 floors, covers a surface area of $2437.761m^2$, and has $517.34m^2$ of floorspace. Scans of homes in their entirety provide opportunities for learning about long-range context, which is critical for autonomous navigation and holistic scene understanding. For the ObjectGoal Navigation task, [8] provided a standard split of 61 training scenes, 11 validation scenes, and 18 testing scenes.

3) RoboTHOR: RoboTHOR [11] is one of the AI2-THOR [14] frameworks that offers simulated environments paired with their physical counterparts to systematically explore and overcome the challenges of simulation-to-real transfer. It serves as a platform where researchers worldwide can remotely test their embodied models in the physical world to evaluate their generalization. The primary aim of RoboTHOR is to assess the generalization of models from simulation to the real world. Currently, it contains a training and validation corpus of 75 simulated scenes, with a split of 60 for training and 15 for validation.

4) Habitat Matterport 3D: Habitat Matterport 3D (HM3D) [9] is a large-scale dataset comprising 1,000 building-scale 3D reconstructions of diverse real-world locations. It contains more than 10,600 rooms and covers approximately 1,920 building floors, providing a navigable space of $112.5km^2$. The dataset is densely annotated with semantic labels by HM3DSem [10], which surpasses the scale, quality, and diversity of object annotations in prior datasets, with 142,646 object instance annotations across 216 3D spaces and 3,100 rooms. Habitat 2022 ObjectNav challenge [12] selected 120 scenes from HM3DSem, and the dataset was divided into 80/20/20 for train/val/test splits, respectively.

TABLE I
SUMMARY OF PUBLIC DATASETS.
* THE DATA OF "NAVIGABLE SPACES" DIRECTLY FROM HM3D [9].
- GIBSON [7] DOES NOT PROVIDE SEMANTIC INFORMATION

Datasets	**Gibson**	**MP3D**	**RoboTHOR**	**HM3D**
Num of Scenes	571	90	75	1000
Navigable Spaces(m^2)*	81.84k	30.22k	0.75k	112.50k
Floors	1400	235	∼75	1920
Instance Annotations	-	50,811	731	142,646

TABLE II
SUMMARY OF CHALLENGES.

Challenge	**Habitat ObjectNav Challenge**			**RoboTHOR Challenge**
	2020	2021	2022	2021
Episodes	∼2803k	∼2803k	∼4002k	∼112k
Num of Targets	21	21	6	12
Complexity	17.09	17.09	13.31	2.06
Best SR	21.08	37.6	68	65.15
Best SPL	8.51	15.6	37	28.84

We summarize the attributes of the above datasets in the Table I. We can see that HM3D has the largest scale and highest quality among them.

B. Online Challenge Benchmarks

Apart from the public datasets, the Embodied AI research community has also organized a series of challenges in recent years, which provide platforms for fair comparisons and attract research attention. We review the two main challenges [5], [11] as follows.

1) RoboTHOR Challenge: The 2021 RoboTHOR Challenge [11] is a continuation of our 2020 RoboTHOR Challenge, held in conjunction with the CVPR Embodied AI Workshop. The challenge focused on the problem of simulation-to-real transfer aiming to encourage researchers to address the problem of simulation-to-real transfer. The RoboTHOR Challenge 2021 deals with the task of Visual Semantic Navigation from ego-centric RGB-D camera input. The agent starts from a random location in an apartment and is expected to navigate towards an object that is specified by its type.

2) Habitat ObjectNav Challenge: The Habitat Object Navigation Challenge [5], [12] has been organized at the CVPR Embodied AI workshop for four years. The objective of this challenge is to develop and evaluate intelligent agents capable of navigating to a specific target object within a previously unseen 3D environment. Participants are provided with a simulation platform, Habitat-Sim [2], which allows them to train and test their algorithms in realistic settings.

We summarize the attributes of the above challenges in the Table II. We can see that the best performance of the same challenge in recent years is consistently improving indicating that Object Goal Navigation is still full of vitality and worthy of researchers' further exploration.

III. Metrics

There are four main metrics for Object Goal Navigation task to evaluate the performance of the agent.

- Success Rate (SR). It is the % of episodes in which the agent stops within a specific distance threshold from an instance of target category:

$$SR = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{1}$$

Where S_i is a binary indicator of success in episode i. An episode is considered successful if, upon calling the STOP action, the agent is within 1.0m Euclidean distance of any instance of the target object category, and the object is visible to an oracle from that stopping position by turning the agent or looking up/down. SR is a positive metric that measures the agent's ability to locate the target object.

- Success weighted by inverse Path Length (SPL) [2]. Obviously, agents that take different paths to navigate to the object will get the same success rate with different efficiency. SPL can tell the difference of them:

$$SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)} \tag{2}$$

where N is the number of episodes, l_i is length of shortest path between goal and target object for episode i and p_i is the length of path taken by agent in episode i. SPL metric accurately measures the efficiency of paths although they all get 100% SR. The SPL metric, however, does not consider the cost of rotating or turning in place. This means that an agent could rotate 360 degrees before moving to obtain more information, yet still achieve the same SPL.

- Distance To the Success(DTS). It describes the distance of the agent from the success threshold boundary when the episode ends:

$$DTS = \max(||x_T - G||_2 - d_s, 0) \tag{3}$$

where $||x_T - G||_2$ is the distance between the agent and the target location at the end of the episode and d_s stands for the success threshold. DTS mainly focuses on those episodes in which the agent failed to navigate to the target object, from which we can tell if the agent came close to succeeding despite the failure.

- Soft SPL. It is a variation of SPL that measures the efficiency of the path taken by the agent, and it considers both successful and failed episodes. Unlike SPL, which is limited to successful episodes, Soft SPL is calculated for all episodes to evaluate the overall efficiency of the agent's trajectory. This metric was first introduced in the Habitat 2020 PointNav challenge [15].

The Habitat Challenge selects SPL as its primary evaluation metric to rank the leaderboard. In cases where there are statistically insignificant differences in SPL, the challenge organizers reserve the right to use additional metrics to determine the winners.

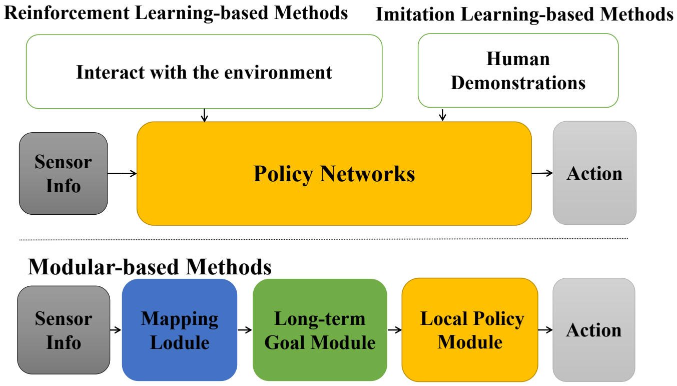

Fig. 2. The three main categories of methods employed in Object Goal Navigation: Reinforcement Learning-based Methods, Imitation Learning-based Methods, and Modular Methods.

IV. Methods

To achieve the task described in Section I, various approaches have been proposed and can be classified into three types: reinforcement learning (RL)-based methods, imitation learning (IL)-based methods, and modular-based methods. End-to-end methods, which include RL and IL methods, face numerous challenges such as high computational costs, and high sensitivity to reward function settings. Conversely, modular-based methods usually have much lower computational costs and higher interpretability, but they heavily depend on the task-specific design of their structures. The general frameworks of these three types of methods are illustrated in Fig. 2.

A. Reinforcement Learning-based Methods

Reinforcement learning (RL) was one of the earliest methods used in the Object Goal Navigation task. In this approach, the agent receives information from sensors (RGB-D, pose) and the target object label, and decides on the next action to take by extracting features from the available information and feeding them into a policy network.

Several RL-based methods have utilized end-to-end training. For instance, the baseline method in the Habitat Challenge, as described in [16], trained a model using DDPPO on multiple nodes. Other methods have proposed graph structures or utilized graphs to help the agent acquire a better understanding of the environment layouts [17]–[20]. [17] introduced a knowledge graph and a context matrix to leverage the inherent relationship between target objects and contextual objects occurring in their surroundings. The method generated a joint embedding which was sent to an LSTM [21] unit and then fed to an A3C model and the authors in [18] introduced a hierarchical object-to-zone (HOZ) graph for scene categorization and goal zone determination. They employed a zone-to-action LSTM to generate actions. Several methods [22]–[24] have also applied transformer [25] encoder and decoder which can deal with time sequences efficiently to address the navigation task. These Attention-based methods also perform well, such as SMTSC [23], which first generates joint features stored in a memory module and then extracts a distribution on possible actions through an attention-based policy network. Some approaches [26], [27] have focused on combining CLIP [28], a strong visual and language model, with RL-based methods. The authors in [27] concentrated on zero-shot tasks. During training, they obtained goal embeddings using a CLIP's visual encoder in ImageNav, while during evaluation, they utilized a corresponding CLIP's textual encoder in ObjectNav. They then sampled actions using a policy network, considering the goal embeddings and other common information. Since the Object Goal Navigation task requires the agent to make decisions using both visual (RGB-D camera) and language (target object) information, these approaches have achieved good performance, especially in the zero-shot domain. Other methods [29]–[31] have improved the task's performance through various methods. For instance, [29] finetuned a simple neural architecture (CNN+RNN) on a large dataset, while [30] created several auxiliary tasks to guide the agent.

B. Imitation Learning-based Methods

Reinforcement learning is often hindered by the challenge of reward engineering. In contrast, methods based on imitation learning [32] can perform better when the training data is sufficiently abundant. Among all published methods, those based on imitation learning have achieved state-of-the-art (SOAT) or near-SOAT results on the Object Goal Navigation task.

Habitat-web [33] provided a dataset containing 80,000 human demonstrations collected through their website, and all existing imitation learning-based methods [33]–[35] depend on it. Habitat-web [33] encoded RGBD information, target label, GPS and compass into a concatenated feature with ResNet and a Semantic Predictor and got actions from a GRU module. This simple architecture made them won the second place in 2021 Habitat Challenge [5]. Nevertheless, imitation learning suffers from poor generalization to new states and difficulty in collecting training data. To address these drawbacks, PIRL-Nav [35] was proposed, which combines RL and IL. PIRLNav fine-tunes policies pre-trained by imitation learning using the PPO algorithm and achieves SOAT performance on MP3D among published methods.

C. Modular-based Methods

Modular-based methods are popular due to their interpretability, efficiency in data collection, and low computational cost, making them a competitive choice for Object Goal Navigation tasks. Most modular methods are map-based [36]–[41], and almost all of them inherit the semantic mapping module from [38] , in which the authors proposed a modular architecture containing semantic mapping, goal-oriented semantic policy and deterministic local policy. PONI [37] attempted to find the target object by predicting a potential function score using a U-net [42] structure on the frontiers and collected training data using MapNet [43]. In contrast, PEANUT [36] employed a PSPNet [44] to directly predict the location of the target object using data collected by the mapping module from [38]. It achieved superior performance compared to PONI due to the significant variation of frontiers based on unseen obstacles, making accurate prediction challenging. Importantly, [41] proposed a robust baseline for exploring the environment using a deterministic policy. They employed four methods to prevent the agent from getting trapped, with the long-term goal being the selection of one of the four corners in a local map. This approach yielded remarkably good results in exploring the environment and searching the target instance. 3D-Aware [39] emphasized the significance of 3D information in navigation tasks. It achieved this by predicting goal corners in a similar manner to [41] but in a learnable manner. Additionally, it identified the target object by utilizing 3D points from various views and [40] attempted to predict the possibility of the target object on each grid in the map using a propability map. In addition, MM1-POMP+ [45] focused on the problem of learning an optimal policy online and proposed POMCP, which allows for training-free online policy learning in unknown environments.

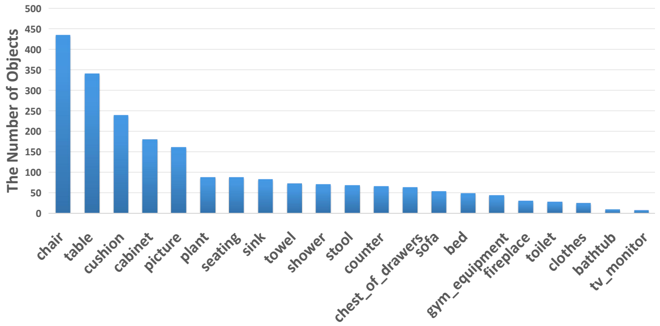

Fig. 3. The distribution of objects in 2200 val episodes of MP3D.

In conclusion, both Modular-based and RL-IL hybrid methods hold significant potential for development in Object Goal Navigation. However, modular-based methods often rely on map-based structures that are vulnerable to influence from detectors such as Mask RCNN [46]. On the other hand, RL-based methods typically suffer from high computational costs, poor sample efficiency, and poor generalization. Collecting or expanding training data of high quality is a challenging task for IL-based methods, and incorrect data can significantly impact performance. Hybrid methods that leverage the strengths of both modular and RL-IL methods have the potential to significantly improve performance on this task.

V. Future research directions

In this section, we will discuss two particularly noteworthy directions that have the potential to advance the field significantly: addressing the long-tail problem and integrating large AI models. By delving into these areas, researchers can push the boundaries of current methodologies and contribute to a more comprehensive understanding of Embodied AI tasks.

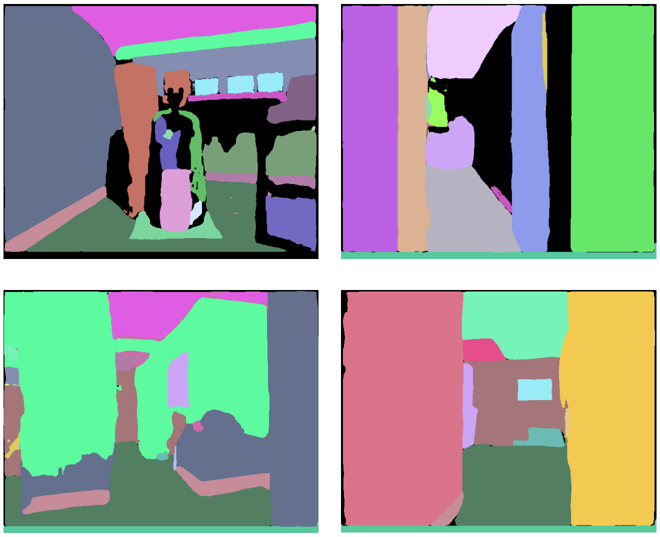

Fig. 4. These are the results of SAM segmentation on first-view images from Habitat-sim.

A. Long-tail Problem

Deep long-tail learning is a crucial topic in computer vision, focusing on the label distribution imbalance in datasets. It has attracted significant research attention [47]. Fig. 3 shows the distribution of target objects in the evaluation episodes of MP3D. Taking into account previous studies' findings [40] and the distribution of semantic annotations in MP3D [8], it can be inferred that the agent performs poorly on objects with few instances. Thus, addressing the long-tail problem is crucial for solving the Object Goal Navigation problem. Considering the application of methods used for long-tail problems such as re-sampling, re-weighting, and metric learning within the context of Object Goal Navigation can be a promising future work. Developing an agent capable of navigating to any object in an unfamiliar house is the ultimate goal of ObjectNav, and there is still a significant amount of crucial and challenging work for researchers to undertake.

B. ObjectNav Meets Large Models

Visual perception plays a pivotal role in the task of Object Goal Navigation, and prior common knowledge can offer the agent general guidance regarding the overall direction, for example chairs are commonly found around tables. Inspired by the recent surge in large-scale visual and language models [48]–[50], it is natural to leverage the advantages of large-scale models to enhance the ability of perceiving the environment and analyzing the instructions. We believe that integrating these large models into the Object Goal Navigation task can significantly enhance an agent's navigation performance.

To demonstrate the potential application of large-scale models in ObjectNav, we present a series of preliminary experimental results utilizing these models. Results in Fig. 4 illustrate the ability of large vision model [49] to provide comprehensive semantic information, thereby enhancing the agent's perception in certain scenes without requiring fine-tuning. Furthermore, large language models like GPT [48] can be employed by providing appropriate prompts. For instance, by giving the prompt "*where is a chair typically located in a house? give the locations in the string format*" to Chat-GPT, we can receive "*dining room, living room, bedrooms, home office, outdoor spaces*" which provide the agent with a rough direction. In conclusion, large-scale models, with their capacity to handle substantial amounts of training data, possess abundant common sense and strong visual perception. As a result, they contribute to assisting agents in their decision-making processes, thereby providing a promising research direction.

VI. Conclusion

In this survey paper, we present a comprehensive review of Object Goal Navigation. We initially discuss public datasets and online challenge benchmarks. Subsequently, we summarize and analyze all evaluation metrics. We also provide an in-depth review of recently proposed methods, dividing them into three categories: Reinforcement Learning-based Methods, Imitation Learning-based Methods, and Modular-based Methods. Finally, we offer our insights on future research directions, including the long-tail problem and the integration of large AI models.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 62103236.

References

[1] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022.

[2] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[3] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.

[4] J. Ye, D. Batra, E. Wijmans, and A. Das, "Auxiliary tasks speed up learning point goal navigation," in *Conference on Robot Learning*. PMLR, 2021, pp. 498–516.

[5] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[6] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3674–3683.

[7] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9068–9079.

[8] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[9] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv preprint arXiv:2109.08238*, 2021.

[10] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva *et al.*, "Habitat-matterport 3d semantics dataset," *arXiv preprint arXiv:2210.05633*, 2022.

[11] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford *et al.*, "Robothor: An open simulation-to-real embodied ai platform," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3164–3174.

[12] K. Yadav, S. K. Ramakrishnan, J. Turner, A. Gokaslan, O. Maksymets, R. Jain, R. Ramrakhya, A. X. Chang, A. Clegg, M. Savva *et al.*, "Habitat challenge 2022," 2022.

[13] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5664–5673.

[14] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu *et al.*, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[15] A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, and D. Batra, "Sim2real predictivity: Does evaluation in simulation predict real-world performance?" *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6670–6677, 2020.

[16] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," *arXiv preprint arXiv:1911.00357*, 2019.

[17] A. Pal, Y. Qiu, and H. Christensen, "Learning hierarchical relationships for object-goal navigation," in *Conference on Robot Learning*. PMLR, 2021, pp. 517–528.

[18] S. Zhang, X. Song, Y. Bai, W. Li, Y. Chu, and S. Jiang, "Hierarchical object-to-zone graph for object navigation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 130–15 140.

[19] F. Zhu, X. Liang, Y. Zhu, Q. Yu, X. Chang, and X. Liang, "Soon: Scenario oriented object navigation with graph-based exploration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 689–12 699.

[20] Y. Lyu, Y. Shi, and X. Zhang, "Improving target-driven visual navigation with attention on 3d spatial relationships," *Neural Processing Letters*, vol. 54, no. 5, pp. 3979–3998, 2022.

[21] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[22] R. Fukushima, K. Ota, A. Kanezaki, Y. Sasaki, and Y. Yoshiyasu, "Object memory transformer for object goal navigation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 288–11 294.

[23] T. Campari, P. Eccher, L. Serafini, and L. Ballan, "Exploiting scene-specific features for object goal navigation," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 406–421.

[24] H. Du, X. Yu, and L. Zheng, "Vtnet: Visual transformer network for object goal navigation," *arXiv preprint arXiv:2105.09447*, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Khandelwal, L. Weihs, R. Mottaghi, and A. Kembhavi, "Simple but effective: Clip embeddings for embodied ai," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 829–14 838.

[27] Q. Zhao, L. Zhang, B. He, H. Qiao, and Z. Liu, "Zero-shot object goal visual navigation," *arXiv preprint arXiv:2206.07423*, 2022.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[29] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi *et al.*, "Procthor: Large-scale embodied ai using procedural generation," *arXiv preprint arXiv:2206.06994*, 2022.

[30] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary tasks and exploration enable objectgoal navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 117–16 126.

[31] O. Maksymets, V. Cartillier, A. Gokaslan, E. Wijmans, W. Galuba, S. Lee, and D. Batra, "Thda: Treasure hunt data augmentation for semantic navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 374–15 383.

[32] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[33] R. Ramrakhya, E. Undersander, D. Batra, and A. Das, "Habitat-web: Learning embodied object-search strategies from human demonstrations at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5173–5183.

[34] K. Yadav, R. Ramrakhya, A. Majumdar, V.-P. Berges, S. Kuhar, D. Batra, A. Baevski, and O. Maksymets, "Offline visual representation learning for embodied navigation," in *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2022.

[35] R. Ramrakhya, D. Batra, E. Wijmans, and A. Das, "Pirlnav: Pretraining with imitation and rl finetuning for objectnav," *arXiv preprint arXiv:2301.07302*, 2023.

[36] A. J. Zhai and S. Wang, "Peanut: Predicting and navigating to unseen targets," *arXiv preprint arXiv:2212.02497*, 2022.

[37] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 890–18 900.

[38] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4247–4258, 2020.

[39] J. Zhang, L. Dai, F. Meng, Q. Fan, X. Chen, K. Xu, and H. Wang, "3d-aware object goal navigation via simultaneous exploration and identification," *arXiv preprint arXiv:2212.00338*, 2022.

[40] M. Zhu, B. Zhao, and T. Kong, "Navigating to objects in unseen environments by distance prediction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 571–10 578.

[41] H. Luo, A. Yue, Z.-W. Hong, and P. Agrawal, "Stubborn: A strong baseline for indoor object navigation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3287–3293.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[43] V. Cartillier, Z. Ren, N. Jain, S. Lee, I. Essa, and D. Batra, "Semantic mapnet: Building allocentric semantic maps and representations from egocentric views," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 964–972.

[44] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[45] F. Giuliari, A. Castellini, R. Berra, A. Del Bue, A. Farinelli, M. Cristani, F. Setti, and Y. Wang, "Pomp++: Pomcp-based active visual search in unknown indoor environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1523–1530.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[47] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[48] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[49] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[50] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.