CVPR #11

Diagnosing Hallucination Problem in Object Navigation

Anonymous CVPR submission

Paper ID 11

Abstract

This work investigates the hallucination problem in object navigation, which leads agents to make incorrect navigation decisions. We identify two kinds of hallucinations: visual grounding and navigation policy. Visual grounding hallucinations are grounding errors from a grounding model that can mislead the agent policy. Policy hallucinations cause the agent to make mistakes even with accurate visual grounding. We analyze how these hallucinations contribute to navigation errors and affect navigation performance, and find that hallucinations about goal objects are the main bottleneck. Finally, we explore the usage of factors like grounding confidence to identify potential directions to mitigate hallucinations in object navigation.

1. Introduction

An embodied agent that is able to navigate to a target object in novel environments has been a long-term goal of embodied AI research [1]. To achieve this, a lot of work has been done to build a better navigation policy [4, 10, 13, 15] and improve the agent's visual grounding [9, 11, 13]. However, the performance of the state-of-the-art navigation agent is still far from perfect [4]. In the navigation process, the agent could make incorrect navigation decisions that lead to failures. For example, stopping at an incorrect object or not navigating to a goal object. While these wrong decisions affect navigation performance significantly, no study has been conducted to deeply analyze why these decisions were made and how we can mitigate them.

Incorrect navigation decision-making could be interpreted as the model having an incorrect belief in the existence of the target object. In this work, we analyze the source of incorrect decisions from the **object hallucination** perspective and diagnose the hallucination problems in object navigation. We first define two main hallucination sources as in Fig. 1. The first one is the visual grounding. In navigation, the agent first needs to perform visual grounding and have an understanding of the environment. The hallucination in the grounding input may cause the agent to make incorrect decisions. The second source is the navigation policy, which processes the grounding input and takes sequential actions. The hallucination of navigation policy appears when the agent makes incorrect decisions when the visual grounding is correct, i.e., the policy does not trust the correct object grounding.

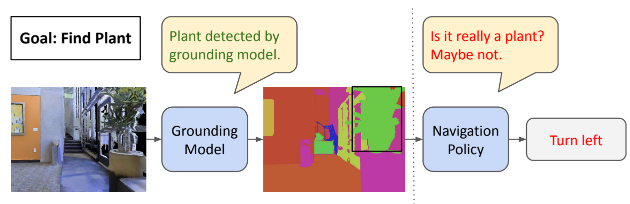

Figure 1. Two main sources of hallucination in object navigation. In the example, although the grounding model predicts correctly, the navigation policy hallucinates and leads to an navigation error.

First, for the grounding input, we show the influence of grounding hallucination by comparing the performance between evaluating with ground truth grounding and predicted grounding. Then, we further investigate the degree of influence from different kinds of objects and how navigation policies leverage the grounding results by providing ground truth for goal or non-goal objects. We then define two major navigation errors and two navigation policy hallucinations and show the correlation between them, from which we conclude that the navigation policy learns to ignore the positive grounding input during imitation learning. Finally, we analyze what grounding factors may help the grounding policies distinguish grounding input hallucinations, such as the area and the confidence of the grounding results, to provide potential directions to mitigate the hallucination problems in object navigation.

In summary, our findings include: (1) The hallucination about goal objects significantly influences navigation performance. (2) The navigation policies can leverage the imperfect grounding of non-goal objects to explore the environment. Only seriously incorrect non-goal object grounding will affect the navigation performance. (3) The navigation policy learns to ignore the positive grounding input during training. (4) In terms of hallucinations on goal objects, we could potentially leverage more detection information, such as grounding confidence, to mitigate them.

2. Related Work

2.1. Object Navigation

There are two main lines of work in object navigation research. The first one is to build a better navigation policy to explore the environment and get closer to the goal object [10, 13, 15]. Among these methods, the learning-based methods train an end-to-end navigation policy with imitation learning or reinforcement learning [4, 10, 13] achieves state-of-the-art results. The other line of research aims to improve the visual perception of the navigation agent to better understand the environment [9–13]. EmbCLIP [9] leverages frozen clip embedding as a more robust visual representation. [11, 12] tried to improve the visual representation of the environment using self-supervised pre-training. Beyond RGB image visual embedding, [10, 13, 15] also leverages more high-level visual information – object detection and semantic segmentation in navigation. For different navigation policies, previous work has shown that imperfect grounding is a major bottleneck for navigation agent [6, 13, 15]. In this work, we try to understand how exactly imperfect grounding affects navigation and how to mitigate this problem.

2.2. Visual Navigation Model Analysis

Prior works in object navigation have used some evaluation methods to understand the navigation models. First, various metrics are proposed to evaluate a navigation episode, such as success rate, distance to goal, SPL, SoftSPL, etc. Then, [3, 13] tried to ablate the semantic segmentation input during evaluation to see how much imperfect segmentation affects the navigation performance. Further, modular navigation methods [5, 6, 15] define different navigation errors to understand where the bottleneck is. However, these evaluations are still relatively superficial, which mainly shows the models' performance but lacks analysis on where and why the performance gap exists. In the vision-and-language navigation (VLN) task, some works performed in-depth model behavior analysis. Zhang et al. [14] tried to diagnose the reason why it is hard for VLN agents to generalize to novel environments. Zhu et al. [16] tried to understand the behavior of VLN agents by designing ablation experiments on the input during evaluation time. We perform detailed analysis for the influence of imperfect grounding in object navigation tasks, including both training and evaluation time, and provide solutions to mitigate the grounding problems.

3. Object Navigation and Hallucination Problems

3.1. Object Navigation

In a typical object navigation task, an agent starts in an unknown environment E with the goal of finding an object from a specific category G, like a chair or cabinet. The agent doesn't know the exact location beforehand. At each step t, the agent receives sensory data O_t. This data typically includes an egocentric RGB-D image, and might also include its position and orientation P_t in some environments. Based on this information, the agent chooses an action a from a set of available actions $\mathcal{A}$, which includes a special "stop" action to indicate it has found the object. The navigation is successful if the agent stops within a certain distance (1 meter) of the target object and can see it without moving.

3.2. Object Navigation Models

In our study, we consider *end-to-end* object navigation models. A typical end-to-end object navigation model first takes the inputs and encodes them into embeddings. For the visual input RGB-D images, there are two kinds of encoders. The first one is a visual encoder like CNN [9], which encodes the RGB-D image into an image feature. The second one is a semantic encoder, which first performs semantic segmentation and then encodes the segmentation results into a visual embedding [10]. Then, the input embeddings are fed into a decision network based on a recurrent neural network or a transformer. In this work, we select a transformer-based architecture [4] that archives state-of-the-art results and follows their training and evaluation setting in experiments. The pipeline of a navigation model is shown in Fig. 1

To better analyze and quantify the influence of grounding, we use semantic-level visual encoders for object navigation models since their embedding is more explainable. For instance, we could acquire the ground truth semantic embedding from the simulator. Specifically, we experiment with two kinds of semantic-level encoders. First, following Ramrakhya et al. [10], we use a Rednet [8] semantic segmentation model trained on in-domain data to predict a semantic segmentation map M_{sem} from the RGB-D image. Then, we use a ResNet [7] to encode the semantic segmentation map into a d dimension embedding. Secondly, following Zhang et al. [14], we calculate the area of each object class from M_{sem} and form a semantic embedding with a dimension of 21 – the number of goal object classes in the MP3D dataset. The value in each dimension is the proportion of pixels that an object occupies in the image. We note these two embedding methods as $\mathrm{Rednet}_{\mathrm{semseg}}$ and $\mathrm{Rednet}_{\mathrm{sememb}}$ respectively in Table 1.

Grounding	Pred		GT All		GT Goal		GT Non-Goal		Shuf Non-Goal	
	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL
Rednet$_{\text{semseg}}$	45.3	15.2	58.0	18.8	57.6	19.1	43.0	14.7	38.0	12.4
Rednet$_{\text{sememb}}$	43.1	13.7	54.9	17.5	56.5	17.6	41.2	13.6	36.6	10.5

Table 1. Comparison of object navigation performance with different ground truth and predicted grounding information provided on different semantic embeddings.

Grounding	PoHall 1	Error 1	PoHall 2	Error 2
Rednet$_{\text{semseg}}$	14.8	38.4	0.03	0.36
Rednet$_{\text{sememb}}$	15.0	36.2	0.03	0.33

Table 2. Quantitative evaluation of two kinds of navigation errors and policy hallucinations in object navigation. The number indicates the average number of hallucinations or errors per episode.

3.3. Hallucinations in Object Navigation

We define two sources of navigation hallucinations to better diagnose the hallucination errors in navigation. The first one is the grounding input hallucinations. In the context of semantic segmentation or object detection, this hallucination can be reflected by grounding metrics like Intersection over Union (IoU). The second one is the hallucinations from the navigation policy. It happens when the grounding input is correct while the navigation policy still makes incorrect decisions. For example, even when the visual grounding part successfully captures the goal object, the agent could still make the wrong decision not to navigate to the detected object.

4. Diagnose Hallucinations in Navigation

4.1. Dataset and Metrics

We use MP3D [2] object navigation dataset for training and evaluation in our experiments. We use imitation learning for model training with the imitation learning dataset collected by Ramrakhya et al. [10], which contains 60k [1] trajectories in 56 training environments with 21 goal object categories. We report the evaluation results on the validation split, containing 2195 episodes in 11 unseen validation environments. For evaluation metrics, we use Success Rate (SR) and Success rate weighted by Path Length (SPL) [1].

4.2. Grounding Hallucination in Object Navigation

How is navigation success affected by grounding hallucination? First, to show the influence of grounding hallucination, in Table. 1, we compare the models trained with Rednet predicted semantic segmentation and test with predicted (Pred) or ground truth (GT All) semantic segmentation. We find that, for both semantic encoding methods, testing with ground truth semantic segmentation improves the navigation performance significantly. This shows that grounding hallucination strongly affects navigation performance.

[1] We exclude the training episodes where the goal object does not belong to the 21 goal objects.

Is the goal object grounding the only grounding feature that matters? To better understand how grounding hallucinations from Rednet affect navigation performance, we break down the influence of grounding hallucinations in different object categories. During the evaluation time, we provide ground truth segmentation of goal objects (GT Goal) and non-goal objects (GT Non-Goal) to the navigation policy. Surprisingly, we find that providing the ground truth grounding of the goal object achieves a similar performance to providing all the ground truth grounding. This shows that *better utilizing the grounding information of goal object is the main bottleneck for the navigation agent.*

We can also observe that providing ground truth non-goal objects does not improve the navigation performance. This raises the question of whether the grounding information of non-goal objects is not essential for navigation. To answer this, we randomly shuffle the 20 non-goal object categories in the semantic embedding during evaluation time. In this case, the grounding information of other objects will be totally incorrect, e.g. a table will become a sofa. For consistency, we keep the shuffle order the same for each step within one episode. From Table. 1, we observe that shuffling non-goal objects (Shuf Non-Goal) decreases navigation performance by a large margin – 6.9% in success rate. This shows that the navigation policy suffers from serious hallucinations of non-goal object grounding. However, imperfect grounding information for non-goal objects from Rednet can already benefit navigation decision-making as well as ground truth information. Therefore, *it is not a significant grounding bottleneck for navigation.*

4.3. Policy Hallucination in Object Navigation

In the last section, we showed that the grounding hallucination of goal objects is the main bottleneck of navigation performance. In this section, we will further investigate the policy hallucination in terms of goal objects.

Grounding	Area	Conf
$\text{Rednet}_{\text{semseg}}$	0.62	0.67
$\text{Rednet}_{\text{sememb}}$	0.61	0.67

Table 3. Results of using the grounding area or confidence to judge the navigation success using a naive Bayesian classifier. The number reported is classification accuracy.

We first define two major navigation errors in object navigation: ignoring the goal object (**Error 1**), which means the goal object appears, but the agent did not navigate to it, and stopping incorrectly (**Error 2**), which means that the agent decides to stop in a place not within 1 meter of a goal object. To remove the influence of grounding hallucination, we then define two kinds of **policy hallucinations** about the goal object. The first one is that when the goal object is correctly detected by the grounding model, the policy does not choose to navigate to it and stop (**PoHall 1**). The second one is that the policy decides to stop when the goal object is not detected by the grounding model (**PoHall 2**). We quantitatively calculate these two policy hallucinations in the following ways. For the first one, we calculate the frequency when the goal object appears and is correctly detected by the grounding model (IoU is larger than 0.1), and the agent didn't successfully stop within 40 steps. For the second one, we calculate the frequency when the agent decides to stop in an incorrect location when no goal object is detected within the last 5 steps of navigation. To compare, we also count the number of two navigation errors during navigation.

The results are shown in Table. 2, we find that PoHall 1 happens frequently and contributes more to navigation Error 1. *This means that the policy will usually ignore the correct grounding input and not navigate to it.* Meanwhile, since PoHall 2 appears only less than 10% of times when Error 2 occurs, when the policy decides to stop, typically, a goal object is detected, whether correct or not, within the last 5 steps. This could be because, during imitation learning, the human labeler sometimes did not see the goal objects, or the Rednet model made hallucinations of false positive predictions, resulting in the demonstration not navigating to a detected goal object. Therefore, the navigation policy learns to ignore some positive grounding results. When the human demonstration stops at the goal object, the grounding model can usually make correct predictions. Therefore, the policy is less likely to stop when there is no goal detected. On the other hand, Error 2 is mainly due to grounding input hallucinations. Since most of the time, when the agent stops incorrectly, a goal object is detected, leading to the wrong decision.

4.4. Mitigating Hallucinations in Object Navigation

After learning more about hallucinations in navigation, we now investigate how we can potentially mitigate them. We look at the second kind of navigation error (**Error 2**), which is the most serious error since it directly causes navigation failure. We already know that Error 2 is mainly caused by grounding input hallucinations. Although grounding hallucinations are inevitable, enabling navigation policies to distinguish these grounding hallucinations and make correct navigation decisions can reduce these navigation errors. To investigate how we can improve the policy network on this, we calculate two key grounding features, grounding confidence and grounding areas, when Error 2 occurs, and the opposite of it occurs – the agent stops successfully. To show whether these features are helpful, we use a naive Bayesian classifier to take their values as input and predict whether the episode is successful or not:

$$P(S_k|f) = \frac{P(f|S_k)P(S_k)}{P(f)} \tag{1}$$

Where S_k indicates episode success or not, and the prior $P(S_k)$ is set to a uniform distribution. f is the average grounding confidence or area in the last 5 steps. We collect the data from all the validation trajectories and randomly split them into training and evaluation sets for the naive Bayesian classifier.

The results are shown in Table. 3, we find that the classification accuracy using grounding confidences is significantly higher than using grounding areas as features. This could be because the navigation area is known by the agent and is one of the reasons for the decision to stop by the agent policy. We also noticed that the classification accuracy using either feature is significantly higher than that of a random guess – 50%. Therefore, *grounding features like confidence that are not currently utilized by the navigation agent could be helpful in mitigating grounding hallucinations.*

5. Conclusion and Discussion

In this work, we study the hallucination problem in object navigation, where the agent has incorrect beliefs about objects. We define the two sources of navigation hallucination and quantitatively analyze their contributions to navigation errors and their influence on navigation performance. For the most critical navigation error, we analyzed the main cause and proposed potential solutions. We hope this work can help the research community understand the hallucination problems in object navigation and provide insights on mitigating them. The limitation is that our analysis focuses on the hallucination problem for end-to-end object navigation models. Therefore, the conclusions may not be generalized to those modular-based navigation models that leverage explicit semantic mapping.

References

[1] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. In *arXiv:2006.13171*, 2020. 1, 3

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 3

[3] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *In Neural Information Processing Systems*, 2020. 2

[4] Shizhe Chen, Thomas Chabal, Ivan Laptev, and Cordelia Schmid. Object goal navigation with recursive implicit maps. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7089–7096. IEEE, 2023. 1, 2

[5] 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. *CVPR*, 2023. 2

[6] Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 8(79):eadf6991, 2023. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[8] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 2

[9] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[10] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek 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*, pages 5173–5183, 2022. 1, 2, 3

[11] Karmesh Yadav, Arjun Majumdar, Ram Ramrakhya, Naoki Yokoyama, Alexei Baevski, Zsolt Kira, Oleksandr Maksymets, and Dhruv Batra. Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav. *arXiv preprint arXiv:2303.07798*, 2023. 1, 2

[12] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. 2

[13] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectgoal navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16117–16126, 2021. 1, 2

[14] Yubo Zhang, Hao Tan, and Mohit Bansal. Diagnosing the environment bias in vision-and-language navigation. *arXiv preprint arXiv:2005.03086*, 2020. 2

[15] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR, 2023. 1, 2

[16] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. *arXiv preprint arXiv:2103.16561*, 2021. 2