Attention to Action: Leveraging Attention for Object Navigation

Shi Chen
chen4595@umn.edu
Qi Zhao
qzhao@cs.umn.edu

Department of Computer Science and Engineering
University of Minnesota

Abstract

Navigation towards different objects is prevalent in daily lives. State-of-the-art embodied vision methods accomplish the task by implicitly learning the relationship between perception and action or optimizing them with separate objectives. While effective in some cases, they have not yet developed (1) a tight integration of perception and action, and (2) the capability to address visual variance that is significant in the moving and embodied setting. To close these research gaps, we introduce a new attention mechanism, which represents the pursuit of visual information that highlights the potential directions of final targets. Instead of working conventionally as a weighted map for aggregating visual features, the new attention is defined as a compact intermediate state connecting visual observations and action. It is explicitly coupled with action to enable a joint optimization through a consistent action space, and also plays an importance role in learning features more robust against visual variance. Our experiments show significant improvements in navigation across various types of unseen environments with known and unknown semantics. Ablation analyses indicate that the proposed method correlates attention patterns with the directions of action, and overcomes visual variance by distilling useful information from visual observations into attention distribution. Our code is publicly available at https://github.com/szzexpoi/ana.

1 Introduction

One of the fundamental goals in artificial intelligence is to develop intelligent agents that can efficiently perceive information from diverse environments, and navigate to different targets with autonomy and adaptability. While humans have little difficulty accomplishing the task, there are two key challenges remaining largely unsolved for embodied agents: First, successful navigation to targets requires close cooperation between perception and action. Existing embodied vision methods either implicitly learn their relationship through repeated trials [6, 9, 34, 36, 38] or leverage separate objectives [20, 24, 33] to independently optimize perception and action, without modeling how they collaborate with each other. Second, existing methods show limited capability in navigating in new environments of which they do not have prior knowledge. Compared to static and passive scenarios in conventional vision tasks, in this moving and embodied context, there is a more significant need for the agents to be able to parse new and diverse visual information and take actions accordingly.

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

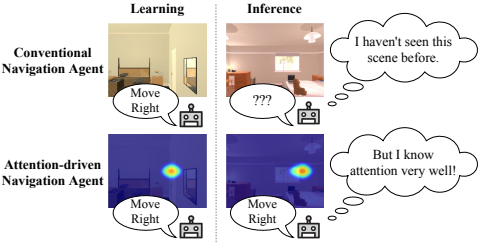

Figure 1: Attention plays an important role in improving the generalizability and interpretability of embodied agents.

As a step towards embodied agents that can perceive and navigate across diverse environments with increased generalizability and interpretability, we borrow inspirations from the evolutionary significance of attention in action [26] and leverage attention as an important interface to integrate perception and action. We define a new attention mechanism that suits the embodied setting and addresses the aforementioned challenges. It differentiates itself in three aspects: (1) Unlike the conventional attention that works as a weighted map for highlighting important regions in different frames, the new attention represents pursuit of visual information that highlights potential directions of the final target and guides navigation towards the corresponding directions. (2) Our method explicitly couples attention with action through the same action space, enabling a joint optimization of where to look and where to move. (3) Visual variance is common due to changes in environments, and results in significant discrepancies between visual features observed from different scenes. Our method leverages attention as a compact intermediate state bridging visual observations and action, distilling useful information instead of conventionally using the exact visual features, and thus is more reliable against variance. Since the need of perception-action integration and the visual variance are significant in the embodied setting, we consider the new attention mechanism a timely contribution to embodied vision, as well as opening a new avenue in attention research with scenarios involving both perception and action.

Specifically, we focus on the object navigation task that involves navigating to an instance of the target object category in unseen environments. We study how to exploit attention for object navigation and develop a novel attention-driven navigation agent (ANA). Augmented with the proposed attention, our method shows significantly enhanced performance in unseen environments with known or unknown semantics. Besides improving the navigation performance, it also provides an interpretable interface for understanding the underlying decision-making process of the agents during navigation.

In summary, this work makes the following contributions:

1. We introduce a new form of attention for the embodied context. It goes beyond the conventional feature-aggregation paradigm of attention, and is applied to the action space to bridge visual observation and action.

2. Aiming at improved robustness against visual variance, we propose to distill useful information from visual observations into attention distribution and learn more robust features for action planning.

3. Through extensive experiments, we demonstrate enhanced performance and generalizability over previous state-of-the-art in object navigation across various unseen en-

vironments. Ablation studies shed light on the role of attention in object navigation, and the key factors for integrating perception and action.

2 Related Works

Our work is most related to previous efforts on attention modeling and visual navigation.

Visual Attention. Inspired by the human visual system that parses visual scenes with a sequence of eye fixations, visual attention has become an increasingly important components in computer vision models. There is a large body of research on attention modeling in conventional vision [2, 10, 17, 19, 35] and embodied vision [3, 8, 11, 21, 25, 37] tasks. These methods typically concentrate on the role of attention in perception, and use it as a weighted map to selectively aggregate visual features from different regions. While showing usefulness, conventional attention does not consider the collaboration between attention and action, and usually requires case-specific analyses [21, 25] to elaborate their relationship. Going beyond the feature-aggregation paradigm, our new attention is designed as the pursuit of information to highlight the potential directions of final targets, and explicitly coupled with action through an adaptive mapping.

Learning-based Visual Navigation. Recent visual navigation methods typically follow a learning-based paradigm to develop policy for action planning. These include methods that make use of recurrent neural networks [3, 8, 11, 22, 30, 31, 32], structured spatial representations [5, 12, 13, 16, 27] and topological representations [7, 28, 29]. Since navigation towards different objects is prevalent in daily life, several recent studies [6, 9, 24, 34, 36, 38] focus on the object navigation task: Zhu *et al.* [38] propose a deep reinforcement learning framework for object navigation, where a picture of the target object is used as the input. Later on, Wortsman *et al.* [34] substitute the picture with language embedding, and develop a meta-learning approach. Aiming for better generalization, Yang *et al.* [36] and Moghaddam *et al.* [24] utilize knowledge about the semantic relationship between different objects, and augment the agent with Graph Convolutional Networks. Chaplot *et al.* [6] develop a neural SLAM model to guide navigation with semantic information. Druon *et al.* [9] incorporate a context grid for encoding the spatial and semantic relationship between objects. The aforementioned methods either implicitly learn the relationship between perception and action [6, 9, 34, 36, 38], or optimize them with separate objectives [20, 24, 33].

Our method differentiates itself in two key aspects: (1) It emphasizes the tight collaboration between perception and action, and explicitly couples attention with action. By mapping attention to the action space, it jointly learns where to look and where to move. (2) Unlike previous methods that typically ignore the visual variance of different environments, we identify it as a critical issue in tasks with embodied paradigms. To overcome variance and improve generalizability, our method utilizes attention as an intermediate state bridging visual observations and action, and determines action based on the attention distribution instead of raw visual features.

3 Attention-driven Navigation Agent

Navigation towards different objects requires understanding visual environments and generating action planning accordingly. Existing object navigation methods [6, 9, 23, 24, 34, 36] pay little attention to the integration of perception and action as well as the visual variance of

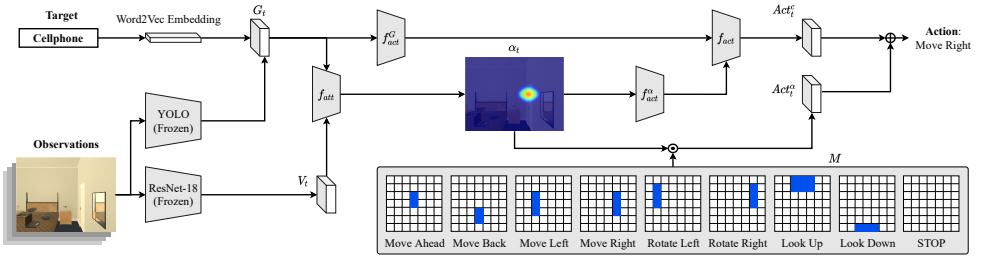

Figure 2: Illustration of the proposed method. The left and the upper part visualize the model architecture, while initialization of the spatial masks M is shown at the bottom right. $\odot$ denotes the dot product while $\oplus$ represents an element-wise summation. For the spatial masks, locations with value 1 and 0 are highlighted in blue and white color, respectively.

diverse environments. As a result, they fall short of navigating in environments where they do not have prior knowledge and have limited generalizability. Attention is an important mechanism in humans that affects where people look to acquire information and where they move based on the observations [26]. In this paper, we propose a new attention to addresses the perception-action integration and visual variance challenges in object navigation.

Our method consists of two principal components: (1) It leverages attention as an intermediate state and learns features from attention distribution for action planning. Compared to raw visual features, attention, as a compact feature representation with well-defined structure (*i.e.* a probabilistic map), is more robust against the visual variance and plays a key role in navigating across diverse environments. (2) Now with the new attention that goes beyond the feature-aggregation paradigm, our method further utilizes an explicit mapping to couple attention with action, and jointly optimizes them through a consistent action space. Besides improving the performance, it also provides an interpretable interface for understanding the underlying rationale behind the agent's decisions.

3.1 Leveraging Attention as an Intermediate State

A key challenge preventing the generalization of embodied agents is the visual variance of diverse environments. The variance leads to significant discrepancies in visual features learned by embodied agents, causing difficulties for them to understand different environments and generate reasonable action planning. To tackle the challenge, our method bypasses the direct use of visual features, and instead uses attention distribution to derive features. As shown in Figure 2, instead of working conventionally as a weighted map, our attention serves as the intermediate state connecting visual observations and action.

Specifically, at each time step t, our method computes the attention map by considering both the visual observations and semantic relationship between different objects. The visual observations encode important knowledge about the environment, while the semantic relationship provide useful priors for object navigation (*e.g.,* coffee mug is likely located near the coffee machine). Following [9], we use ResNet-18 [15] to extract visual features $V_t \in \mathbb{R}^{2048\times14\times14}$ from four recent observations in egocentric view, and represent the semantic relationship with a context grid $G_t \in \mathbb{R}^{1\times16\times16}$ that encodes the cosine similarity between word embeddings of object categories for the target and objects detected in the current frame. The two types of features are first processed independently with convolutional and adap-

tive pooling layers to obtains features with consistent spatial dimension $\hat{V}_t \in \mathbb{R}^{128\times7\times7}$ and $\hat{G}_t \in \mathbb{R}^{16\times7\times7}$, and then concatenated to compute the attention $\alpha_t \in \mathbb{R}^{1\times7\times7}$:

$$\alpha_t = \sigma(f_{att}([\hat{V}_t; \hat{G}_t])) \tag{1}$$

where f_{att} denotes the convolutional layers for computing the additive attention, $[\cdot]$ represents the concatenation of features, and σ is the softmax activation function.

Upon obtaining the attention map, instead of multiplying it with visual features, we directly learn features from the attention distribution. Benefiting from the structure and compact nature of attention, the features learned under this paradigm are more robust against visual variance and enables better generalization across different environments. The action in our approach is determined based on the features derived from both attention distribution and semantic relationship between objects:

$$Act_t^c = f_{act}([f_{act}^{\alpha}(\alpha_t); f_{act}^{G}(G_t)]) \tag{2}$$

where f_{act}^{α} and f_{act}^{G} are convolutional layers that further encode attention and context grid, f_{act} denotes fully-connected layers that derive the action likelihood based on flattened features, and Act_t^c is the unnormalized likelihood of candidate actions.

The aforementioned method distills useful knowledge from raw visual features into attention, addressing the issues of visual variance without losing important information (see supplementary materials for details). More importantly, it also enables an explicit integration of perception and action as detailed in the next subsection.

3.2 Coupling Attention with Action

Moving forward, we propose to explicitly couple attention with action and jointly optimize them through a consistent action space to further improve the navigation performance.

The principal idea behind the method is to learn a set of spatial masks that map attention distribution to different candidate actions. The masks serve as action templates that highlight the locations to be explored after performing different candidate actions, and explicitly model the relationship between attention and action. The agent is more likely to perform an action, if its attention is allocated towards the locations in the associated mask:

$$Act_t^{\alpha} = \sum M \cdot \alpha_t \tag{3}$$

where $M \in \mathbb{R}^{k\times7\times7}$ represents the spatial masks and k is the number of candidate actions. Act_t^{α} denotes the unnormalized likelihood of candidate actions mapped from the attention. The summation is performed on spatial dimensions of the attention map.

We initialize the spatial masks M as binary maps based on the average attention patterns for different actions (Figure 2, bottom right; see Section 4.3 for details). To support generalization across various scenarios, we further optimize the masks with the other components (*e.g.*, f_{act} for deriving action) through interacting with various environments. Therefore, the agent can leverage prior knowledge to efficiently establish the relationship between attention and action, and adaptively refine it in a data-driven manner.

Our method determines its final action Act_t by considering both the features derived from attention and contextual information (*i.e.*, Act_t^c) and the alignment between attention and spatial masks (*i.e.*, Act_t^{α}). We incorporate a trainable balance factor $\beta \in \mathbb{R}$ to adaptively combine the two likelihood of actions:

$$Act_t = \sigma(Act_t^c + \beta \cdot Act_t^{\alpha}) \tag{4}$$

where σ is the softmax activation function for normalizing the likelihood of actions.

By leveraging attention as an intermediate state to learn features and explicitly coupling it with action, our method establishes a new paradigm for using attention to integrate perception and action. Our experiments demonstrate its effectiveness on improving the navigation performance across various environments. Our analyses also shed light on the decision-making process, and highlight the key components for integrating perception and action.

4 Experiments and Analyses

In this section, we compare our method with existing state-of-the-art, and analyze the role of attention in object navigation. We report results that validate the effectiveness of our method (Section 4.2), and perform ablation studies that shed light on two research questions that have yet to be answered: (1) Where do embodied agents look while navigating to objects? (Section 4.3), and (2) What are the keys for integrating perception and action? (Section 4.4). Additional analyses and qualitative results are provided in the supplementary materials.

4.1 Experiment Settings

Dataset. Following recent state-of-the-art methods for object navigation [9, 24, 34, 36, 38], we evaluate our method with the popular AI2-THOR [18] framework. It provides 120 photo-realistic environments with four room types (Living Room, Bathroom, Kitchen and Bedroom), where each room type has 30 different environments. We use 9 candidate actions in our experiments, including *Move Ahead, Move Back, Move Left, Move Right, Rotate Left, Rotate Right, Look Up, Look Down* and *STOP* (see supplementary materials for details).

Evaluation Protocols. We demonstrate the effectiveness and generalizability of our method with two popular evaluation settings used in previous state-of-the-art: (1) **Unseen environments with known semantics**: Following [24, 34], we split 30 environments per room type into 20/5/5 for training/validation/testing. Agents are trained with environments of all types, and target objects (see supplementary materials for details) for training and evaluation are consistent. Therefore, while agents have not seen the environments for evaluation during training, they still have knowledge about their layouts and corresponding targets. For testing, we perform inference for 250 episodes per room type, where the environment, initial state and target are randomly chosen. (2) **Unseen environments with unknown semantics**: We also experiment with a more challenging setting where both room types and target objects for evaluation are unknown during training. Specifically, we follow [9], training agents with environments from two room types (*i.e.,* Living Room and Bathroom, Kitchen and Bedroom) and evaluating them on the others. Target objects for evaluation are determined by selecting objects closest to training targets on the word embedding space. This allows us to evaluate the effectiveness of agents on navigating to semantically relevant objects. To be consistent with [9], only 10 environments per room type are used for training (environments for training/validation/testing are adjusted accordingly), and evaluation is carried out on 250 random episodes per environment/target.

We use two popular evaluation metrics, including Successful Rate (SR) and Success weighted by normalized Path Length (SPL) [1]. An episode is successful if the agent releases a *STOP* signal within 300 steps, and an instance of target object is visible and within 1 meter.

		Living Room		Bathroom		Kitchen		Bedroom		Average	
		SR (%)	SPL (%)	SR (%)	SPL (%)	SR (%)	SPL (%)	SR (%)	SPL (%)	SR (%)	SPL (%)
Known Semantics	A3C [23]	-	-	-	-	-	-	-	-	33.40	14.68
	GCN [34]	-	-	-	-	-	-	-	-	35.13	15.47
	SAVN [34]	21.60	7.71	69.60	28.49	43.60	17.80	29.20	8.65	40.86	16.15
	GVE [24]	25.20	9.41	**75.60**	**31.03**	45.60	17.93	27.60	8.06	43.80	17.27
	Spatial Context [9]	45.23	16.27	59.20	22.89	60.20	22.52	40.60	13.57	51.31	18.84
	ANA	**58.16**	**20.61**	68.08	25.07	**75.36**	**28.72**	**48.64**	**16.58**	**62.56**	**22.75**
Unknown Semantics	A3C [23]	23.52	6.34	16.32	3.36	9.65	3.74	6.00	2.46	13.87	3.97
	GCN [34]	14.52	2.40	11.81	1.84	16.36	5.10	6.11	1.73	12.20	2.77
	Spatial Context [9]	37.03	12.38	49.48	10.28	31.07	4.70	21.63	3.82	34.80	7.80
	ANA	**57.09**	**16.40**	**55.56**	**13.88**	**54.42**	**12.22**	**32.32**	**8.02**	**49.85**	**12.63**

Table 1: Comparison between our model and state-of-the-art in unseen environments with known (top) and unknown (bottom) semantics. Results are averaged across 5 runs.

		Living Room		Bathroom		Kitchen		Bedroom		Average	
		SR (%)	SPL (%)	SR (%)	SPL (%)	SR (%)	SPL (%)	SR (%)	SPL (%)	SR (%)	SPL (%)
Known Semantics	Baseline	45.23	16.27	59.20	22.89	60.20	22.52	40.60	13.57	51.31	18.84
	Baseline w/o spatial masks	43.24	15.22	67.80	21.21	68.87	24.16	35.95	11.02	53.97	17.91
	Baseline w/ attention	44.99	15.19	62.27	21.10	63.18	20.54	45.62	14.50	54.02	17.83
	Baseline w/ coupling	46.80	14.87	65.60	20.72	66.00	21.90	44.40	14.15	55.70	17.91
	ANA (full method)	**58.16**	**20.61**	**68.08**	**25.07**	**75.36**	**28.72**	**48.64**	**16.58**	**62.56**	**22.75**
Unknown Semantics	Baseline	37.03	12.38	49.48	10.28	31.07	4.70	21.63	3.82	34.80	7.80
	Baseline w/o spatial masks	43.39	11.30	38.36	12.72	31.79	8.40	33.04	6.78	36.64	9.80
	Baseline w/ attention	38.52	10.16	43.44	12.92	48.08	10.52	**40.29**	**10.90**	42.58	11.12
	Baseline w/ coupling	32.27	9.26	28.76	9.62	39.52	9.72	33.88	9.70	33.61	9.58
	ANA (full method)	**57.09**	**16.40**	**55.56**	**13.88**	**54.42**	**12.22**	32.32	8.02	**49.85**	**12.63**

Table 2: Comparison between our method and different baselines in unseen environments with known (top) and unknown (bottom) semantics. Results are averaged across 5 runs.

4.2 Results

Comparison with the state-of-the-art. We compare our method with the following state-of-the-art: **A3C** [23] is a strong baseline for object navigation, **SAVN** [34] uses meta-learning to enhance model generalizability, **GCN** [36] and **Spatial Context** [9] (our **Baseline**) take advantage of prior knowledge about object relationship, **GVE** [24] incorporates both meta-learning and prior knowledge. According to Table 1, our attention-driven navigation agent (**ANA**) significantly improves the navigation performance across various environments. In environments with known semantics (top panel), our method outperforms the previous best-performing state-of-the-art (Spatial Context) by 11.2% and 3.9% of absolute gain in average successful rate and SPL. For the more challenging setting with unknown semantics (bottom panel), it leads to even larger improvements in successful rate (15.1%) and SPL (4.8%).

Comparison with the baselines. To evaluate the effectiveness of the proposed components in our method, we further conduct experiments on various baselines of the model. **Baseline w/o spatial masks** incorporates conventional soft attention with the Spatial Context [9] baseline to selectively aggregates visual features, without integrating perception and action with the proposed spatial masks. **Baseline w/ attention** includes the proposed attention mechanism but without mapping attention distribution to the action space. **Baseline w/ coupling** couples attention with action, but replaces the proposed attention with conventional attention. As shown in Table 2, without considering the perception-action integration and visual variance, introducing conventional attention to the Baseline fails to provide a visible improvement. On the contrary, our full method achieves a significant boost of performance over the Baseline, and also considerably outperforms baselines with one proposed components (*i.e.*, attention and coupling). The results demonstrates the advantages and the integral design of the proposed method. It is also noteworthy that the proposed attention alone improves performance dramatically with unknown semantics, showing its effective-

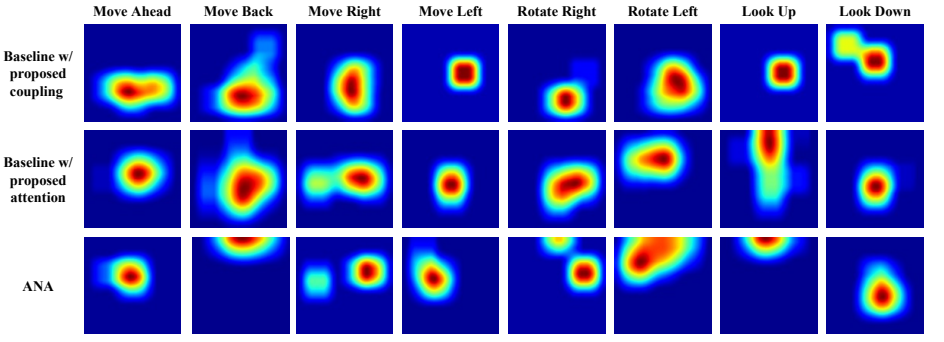

Figure 3: Attention patterns for different candidate actions. Conventional attention for feature aggregation (1st rows) does not have distinct patterns for different actions, while our attention as the intermediate state (2nd row) correlates with the directions of actions. Further coupling attention with action (3rd row) learns more discriminative attention patterns.

ness in distilling useful information to address visual variance, which has been a challenge with unknown semantics. Coupling the proposed attention with action leads to further improvements in both settings.

4.3 Where Do Embodied Agents Look While Navigating to Objects?

A distinct property of the proposed method compared to previous object navigation methods [6, 9, 23, 24, 34, 36] is that it provides an interpretable interface to study the relationship between perception and action. In this section, we report a more in-depth analysis to understand how perception is correlated with action, and where do agents look while navigating. Specifically, we consider three methods (*i.e.*, Baseline w/ proposed coupling, Baseline w/ proposed attention, and ANA), and compare their average attention patterns among different candidate actions. Three key observations can be drawn from the experimental results:

Feature aggregation based attention is not directly correlated with action. Conventional feature aggregation based attention focuses on the role of attention in perception, and pays little attention to its correlation with action. As shown in the 1st row of Figure 3, there is no obvious attention pattern for most of the candidate actions (*e.g., Move Back* and *Rotate Left*). The observation supports our results in Table 2, which shows that coupling conventional attention with action (*i.e.,* Baseline w/ proposed coupling) does not bring reasonable improvements as they are not directly correlated.

Attention as an intermediate state is correlated with action. Unlike the conventional attention, our new attention, as an intermediate state connecting visual observations and action, shows that clear patterns associated with different actions. According to the 2nd row of Figure 3, when performing actions towards the right direction (*i.e., Move Right* and *Rotate Right*), the agent naturally pays focused attention to the right side. Similarly, when adjusting the camera view (*i.e., Look Up* and *Look Down*), it concentrates on the corresponding directions to indicate the pursuit of additional information. The observation validates the effectiveness of our new attention in bridging perception and action.

Coupling attention with action learns discriminative attention. The aforementioned observations validate the advantages of leveraging our attention to close the gap between perception and action, and suggest the potential benefits of explicitly modeling their relation-

	Similarity (SIM)
Conventional Attention	0.023
Living Room	0.312
Bathroom	0.324
Kitchen	0.341
Bedroom	0.305
Obstacle-free	0.337
w/ obstacles	0.306

Table 3: Alignment scores (SIM) between attention and spatial masks.

	Known Semantics		Unknown Semantics	
	SR (%)	SPL (%)	SR (%)	SPL (%)
Random-Uniform	54.81	18.62	48.31	10.46
Random-He	54.95	18.92	49.04	10.33
Fixed Mapping	61.04	20.30	49.60	11.35
Dynamic Mapping	57.94	18.55	42.07	10.97
ANA	**62.56**	**22.75**	**49.85**	**12.63**

Table 4: Comparison between different spatial masks.

ship. Following the inspiration, our full method ANA explicitly couples attention with action for a joint optimization. According to the 3rd row of Figure 3, it learns more discriminative patterns of attention, which also leads to significantly improved navigation performance. For several actions that previously do not have a clear attention pattern (*e.g.*, *Move Left*) or have attention patterns not fully distinguishable from the others (*e.g.*, *Rotate Right*), our method correlates the attention with their moving directions.

4.4 What are the Keys for Integrating Perception and Action?

Analyses in Section 4.3 show that our method tightly couples attention and action, resulting in discriminative attention patterns and enhanced navigation performance. In this subsection, we perform experiments to study its key factors for integrating perception and action:

Cooperation of attention, action and context. Our method explicitly models the relationship between attention and action through trainable spatial masks, and determines action by considering the alignment between attention distribution α_t (see Equation 3) and the spatial masks M. To complement the qualitative results in Figure 3, we adopt the Similarity (SIM) score [9] widely used in attention evaluation, and quantitatively measure the alignment between attention distribution and the spatial mask for action performed at each time step. The higher SIM score, the more likely the agent will look at the moving direction of the corresponding action, thus tighter collaboration between perception and action. Results reported in Table 3 show that our method maintains reasonable SIM scores across different settings, as opposed to the Baseline w/ proposed coupling whose attention (Conventional Attention) is not directly related to action. The observation demonstrates the effectiveness of our method on modeling the relationship between attention and action. We further analyze the SIM scores with respect to different environments, *i.e.*, various room types, and obstacle-free vs. w/ obstacles, which suggests that the attention-action relationship can be environment-dependent, and taking into account the contextual information (*i.e.,* Act_t^c in Equation 4) helps navigation through diverse environments.

Prior knowledge and refinement of relationship. Results in Table 3 show various degrees of alignment between attention and spatial masks. To study the effectiveness of different choices of spatial masks, we compare the proposed method with four alternatives: (1) Randomly initializing the masks with an uniform distribution between 0 and 1 (Random-Uniform); (2) Initializing with method proposed in [14] (Random-He); (3) Utilizing fixed masks to align attention with directions of action (Fixed Mapping); and (4) Dynamically determining the spatial masks (Dynamic Mapping, see supplementary materials for details).

As reported in Table 4, Fixed Mapping significantly outperforms its counterparts with randomly initialized masks, suggesting value with the prior knowledge used for initializing the masks. Dynamic Mapping considers the influences of environments, and achieves slightly better Successful Rate with known semantics than random initialization. The perfor-

mance, however, does not increase in environments with unknown semantics, as the learned attention-action relationship from the training data may not generalize with the absence of prior knowledge. Unlike the compared methods, our full method achieves the best performance by utilizing prior knowledge to establish initial attention-action relationship, and further refining it through interacting with various environments.

5 Conclusion

We introduce ANA, an object navigation method that leverages a novel attention mechanism to integrate perception and action. Unlike existing methods that utilize attention for feature aggregation, our method designs attention as the pursuit of visual information. It distills useful information from visual observations into attention to overcome visual variance, and explicitly couples attention with action to enable a joint optimization of where to look and where to move. Through extensive experiments, we demonstrate the advantages of our method in navigating across various scenarios. Additionally, our analyses highlight the role of attention in object navigation and the key components for integrating perception and action. We hope that this work will be helpful for future development of attention methods for embodied vision, and inspire analyses of the decision-making process.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

[1] Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir Roshan Zamir. On evaluation of embodied navigation agents. *Arxiv*, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019.

[5] Devendra Singh Chaplot, Emilio Parisotto, and Ruslan Salakhutdinov. Active neural localization. In *International Conference on Learning Representations*, 2018.

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

[7] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12872–12881, 2020.

[8] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2018.

[9] Raphael Druon, Yusuke Yoshiyasu, Asako Kanezaki, and Alassane Watt. Visual object search by learning spatial context. *IEEE Robotics and Automation Letters*, 5(2):1279–1286, 2020.

[10] Gamaleldin Elsayed, Simon Kornblith, and Quoc V Le. Saccader: Improving accuracy of hard attention models for vision. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Conference on Neural Information Processing Systems*, volume 32, 2019.

[11] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Conference on Neural Information Processing Systems*, page 3318–3329, 2018.

[12] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4089–4098, 2018.

[13] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7272–7281, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, page 1026–1034, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] Joao F. Henriques and Andrea Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[18] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[19] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3242–3250, 2017.

[20] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. In *International Conference on Learning Representations*, 2019.

[21] Anthony Manchin, Ehsan Abbasnejad, and Anton van den Hengel. Reinforcement learning with attention that works: A self-supervised approach. In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *International Conference on Neural Information Processing*, pages 223–230, 2019.

[22] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J. Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, Dharshan Kumaran, and Raia Hadsell. Learning to navigate in complex environments. In *International Conference on Learning Representations*, 2017.

[23] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, page 1928–1937, 2016.

[24] Mahdi Kazemi Moghaddam, Ehsan Abbasnejad Qi Wu, and Javen Shi. Optimistic agent: Accurate graph-based value estimation for more successful visual navigation. In *IEEE Winter Conference on Applications of Computer Vision*, 2021.

[25] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. In *Conference on Neural Information Processing Systems*, pages 12329–12338, 2019.

[26] Donald A. Norman and Tim Shallice. Attention to action. *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4*, pages 1–18, 1986.

[27] Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. In *International Conference on Learning Representations*, 2018.

[28] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. In *International Conference on Learning Representations*, 2018.

[29] Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphael Marinier, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *International Conference on Learning Representations*, 2019.

[30] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2610–2621, 2019.

[31] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *European Conference on Computer Vision*, pages 38–55, 2018.

[32] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6622–6631, 2019.

[33] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6622–6631, 2019.

[34] Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6743–6752, 2019.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[36] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. In *International Conference on Learning Representations*, 2019.

[37] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl Muller, Jake Whritner, Luxin Zhang, Mary Hayhoe, and Dana Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 6811–6820, 2020.

[38] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE International Conference on Robotics and Automation*, pages 3357–3364, 2017.