Evidential Active Recognition: Intelligent and Prudent Open-World Embodied Perception

Lei Fan[1], Mingfu Liang[1], Yunxuan Li[1], Gang Hua[2] and Ying Wu[1]
[1]Northwestern University [2]Wormpex AI Research
{leifan,mingfuliang2020,yunxuanli2019}@u.northwestern.edu,ganghua@gmail.com,yingwu@northwestern.edu

Abstract

Active recognition enables robots to intelligently explore novel observations, thereby acquiring more information while circumventing undesired viewing conditions. Recent approaches favor learning policies from simulated or collected data, wherein appropriate actions are more frequently selected when the recognition is accurate. However, most recognition modules are developed under the closed-world assumption, which makes them ill-equipped to handle unexpected inputs, such as the absence of the target object in the current observation. To address this issue, we propose treating active recognition as a sequential evidence-gathering process, providing by-step uncertainty quantification and reliable prediction under the evidence combination theory. Additionally, the reward function developed in this paper effectively characterizes the merit of actions when operating in open-world environments. To evaluate the performance, we collect a dataset from an indoor simulator, encompassing various recognition challenges such as distance, occlusion levels, and visibility. Through a series of experiments on recognition and robustness analysis, we demonstrate the necessity of introducing uncertainties to active recognition and the superior performance of the proposed method.

1. Introduction

Passive visual recognition, encompassing a broad range of approaches such as image-based and video-based techniques [17, 27, 37, 56], has experienced tremendous success in recent decades. Nonetheless, a contrasting approach, the active recognition system [1, 2, 4], offers unique advantages. Namely, it benefits from the ability to move within and perceive its environment, allowing it to dynamically determine visual inputs, as opposed to relying on pre-captured images.

A number of approaches [5–7, 32] in active recognition have been proposed over the years with information gain quantification [15], auxiliary task modeling [13, 41], and discriminative feature discovery [29]. Given its sequential nature, prevalent approaches [16, 20, 29, 55] involve combining recognition models, such as object detectors or classifiers, with reinforcement learning techniques. The rewards, which can take the form of recognition accuracy, serve as incentives for the policy and are directly associated with the outputs of the recognition model. Nevertheless, the reward could be misleading when deploying the agent in open environments [25]. Consider an agent being queried with a target in the initial observation. The target may be out of view during subsequent movements, *i.e.*, an unexpected input for the recognition system. As a consequence, the recognition might fail to provide rewards that accurately represent the worth of actions being taken, impeding effective policy learning.

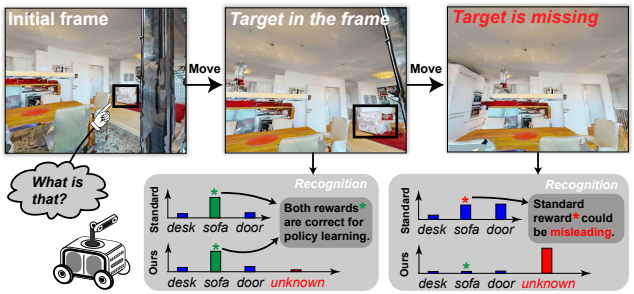

Figure 1. An overview of active recognition agents operating in an open-world environment with different recognition models. The agent is initially queried to perform recognition. In standard practice, the recognition model is unable to identify its own failures, especially when confronted with unexpected inputs, which may lead to incorrect rewards for the policy. In contrast, our approach estimates uncertainty at each step and fuses the collected evidence to provide a final prediction.

Regarding this challenge, our insight is derived from an instinctive understanding of how humans learn from experience. We consciously weigh our learning based on the level of trust we have in our observations. When faced with a high degree of uncertainty, we exercise caution in drawing conclusions from our experiences. Likewise, the agent should evaluate the uncertainty in its observations to avoid being negatively influenced by unexpected inputs, such as when the target object exhibits high ambiguity or is simply not present in the input.

Besides, compared to passive recognition, the importance of modeling uncertainty is more prominent in active recognition, since the operating environment for agents is inherently open and unpredictable. This area, however, remains underexplored as recent methods often resort to recognition models built under the closed-world assumption and then directly using softmax outputs as the confidence measure. In this paper, we choose to model active recognition as a sequential evidence-gathering process, wherein the agent learns to explore trustworthy observations and gives the final prediction based on a collection of opinions. The comparison of the proposed method with standard active recognition methods is depicted in Figure. 1, where our agent can assign high uncertainties to inputs lacking target objects. After actively exploring multiple views, the final prediction of the target is derived from a combination of collected opinions using Dempster-Shafer's combination theory [31].

Furthermore, to evaluate active recognition approaches, we collect a new dataset from an existing indoor simulator [47, 51] where the target could be queried using a bounding box. In addition, each test instance in our dataset is accompanied by a recognition difficulty level, taking into account factors such as visibility, observed pixels, and distance. We expect that the difficulty level will more effectively demonstrate the value of active recognition during evaluation, as the agent can potentially overcome these challenges, unlike in passive recognition.

The primary contributions of this paper are as follows: (1) We develop a novel active recognition approach using evidence theory to address the challenges of open-world environments, which are largely overlooked by other methods. (2) We build a new dataset for evaluating active recognition agents, encompassing 13200 test instances across 27 different indoor object categories. Importantly, each test instance is assigned a recognition difficulty level for a more comprehensive assessment of methods. (3) We conduct extensive experiments and analysis of the proposed method and other baselines. These investigations demonstrate the trustworthiness and superior performance of our method.

2. Related Work

Active recognition. Embodied artificial intelligence, a long-standing field driven by enabling agents to learn through interactions, has been investigated across various streams [12, 23, 33, 39]. As a notable branch, active recognition [1, 4, 8, 13, 22, 25, 36] lets the agent explore observations with its own incentives, thereby achieving improved recognition performance.

Recent approaches [16, 20, 29, 55] in active recognition typically represent the recognition as a Markov Decision Process and rely on Reinforcement Learning (*RL*) to learn moving strategies from interactions. Unlike other *RL* scenarios where the reward function is explicitly specified by the environment [38, 48], the reward function in active recognition is associated with the recognition module, which is defined as information gain [36], the recognition accuracy [13, 29, 30, 55], or the recognition score [4, 16]. In [29, 42], the authors propose an active recognition agent with modules to aggregate historical information and predict the next views. The policy learning is supervised by a binary reward function checking whether the top prediction is correct. In parallel, [16] selects the surrogate reward, defined as the detection score, during policy learning.

However, these reward definitions might not adequately reflect the value of actions when the agent encounters unexpected visual inputs. This is because the recognition model was developed within a closed-world setting. Improper rewards can compromise the recognition policy after training.

Uncertainty estimation. Conventional visual recognition models are unable to identify their own failures. Nonetheless, this ability could be essential when it comes to real-world applications, particularly for embodied agents. Bayesian neural networks [18, 24, 34, 35] equip the recognition models predicting uncertainties as the mutual information between data and parameters. In stark contrast, Evidential Deep Learning [3, 49] established on the Dempster-Shafer Theory has been investigated in different applications [9, 14, 21, 26], which learn the prior of the categorical prediction directly. Inspired by the success of uncertainty quantification of evidential approaches, the proposed method formulates active recognition as a sequential multi-source evidence fusion under the same frame of discernment. With this modeling, we can more effectively combine knowledge in consecutive observations and handle open-world challenges.

3. Task Settings and Notations

We describe our setting by deploying the active object recognition agent in an indoor environment.

The recognition agent is spawned at a location in an indoor environment without prior knowledge about the scene, *e.g.*, the map. At the initial timestep $t = 1$, the agent is queried with a target x in the current visual observation v^1, selected by a bounding box q_x^{box}. And a total of T timesteps is allowed for the agent to obtain the final prediction of the target category. Besides the final step, the agent could take an additional action $a^t \in \mathcal{A}$ at $t = 1, \ldots, T-1$ to change its viewing point, where $\mathcal{A}$ is the action space. By taking movements, the agent could improve its recognition performance by collecting information and give the final predicted label $\hat{y}$ at $t = T$. It is worth noting that the setting could be smoothly generalized to other simulators and different types of targets, like active scene recognition [20, 29, 44, 54].

To provide details about our setup, the observation v is captured by an RGB camera mounted at the height of 1 meter above the ground, featuring a height-by-width resolution of 640×800. The action space is defined as $\mathcal{A} =$

{move_forward, turn_left, turn_right, look_up, look_down}. Actions allow the agent to move 0.25m, turn 10 degrees, or tilt 10 degrees. Previous works on active recognition [16, 55] usually do not contain look_up and look_down in to their action space. This may potentially compromise the agent's performance, particularly in the case of objects situated at higher locations in the observation. As the agent forms multi-round interactions with the environment, the objective of an effective approach lies in three major components, including visual classification, information aggregation, and intelligent moving policies, which we will elaborate on in later sections.

4. Evidential Active Recognition

In this section, we introduce a model called Evidential Active Recognition, which is designed to address the challenges associated with developing agents in an open-world context. We comprehend active recognition as an evidence-collecting procedure in which the agent explores and gathers knowledge to make predictions. Evidence for each class is estimated on a per-step basis, complemented by an additional term that describes the uncertainty. Intuitively, as the agent freely moves in the environment, uncertainty appears when the target is absent, or when the viewpoint is not optimal for acquiring discriminative features. To capture the uncertainty, the Dirichlet prior is placed for known classes [49] to provide trustworthy results. In comparison to active recognition approaches that utilize softmax outputs, the estimated uncertainty then plays a key role in policy learning by redefining the reward function. The final multinomial opinion on the target category follows the Subject Logic [10, 31] to fuse evidence from different views at the last step.

4.1. Preliminaries

Current visual recognition models often rely on the softmax operator to give probabilities. Nonetheless, as the probability is normalized, it could lead to overconfident predictions or even failures when handling unknown inputs. Evidential deep learning approaches [49], on the contrary, choose to develop uncertainties under the scheme towards subjective probabilities [10, 31]. For the next paragraph, we will start by introducing the formulation of evidential deep learning for single-frame prediction.

For a K-class recognition task, the frame of discernment $\Theta = \{k, 1 \leq k \leq K\}$ contains K exclusive singletons, *e.g.*, class labels. Considering the visual observation v^t at timestep t, we measure the mass in mutually exclusive propositions with a belief function b_k^t leveraging the Dempster-Shafer Theory of Evidence (DST) [10]. By providing an overall uncertainty mass u^t, these $K+1$ mass values satisfy

$$\sum_{k=1}^{K} b_k^t + u^t = 1, \quad 0 \leq u^t, b_k^t \leq 1. \tag{1}$$

The evidence e_k^t is defined as the support evidence collected from the current observation v^t to k^{th} singleton. To form the opinion, the evidence e_k^t is introduced to associate the belief function with Dirichlet distribution parameters α_k^t by

$$b_k^t = \frac{e_k^t}{S^t} \text{ and } u^t = \frac{K}{S^t}, \tag{2}$$

where $S^t = \sum_{k=1}^{K} \alpha_k^t = \sum_{k=1}^{K} (e_k^t + 1)$. As the conjugate prior for the multinomial distribution, the Dirichlet distribution characterized by $\alpha^t = [\alpha_1^t, \ldots, \alpha_K^t]$ is used to derive the subject opinion, *i.e.*, the belief function as $b_k^t = e_k^t / S^t$. Accordingly, the overall uncertainty u^t arises if no significant evidence is being collected from known classes.

From the DST, the masses defined in Equation. 1 is a reduction from the general hyper-opinion set $2^\Theta = \{\emptyset, 1, \ldots, K, \{1,2\}, \ldots, \Theta\}$ that contains intimidating 2^K propositions. As no mass is assigned to the empty-set proposition, *i.e.*, $b_\emptyset^t = 0$, the uncertainty u^t is consequently the summation of contained non-singleton belief masses. Namely, the overall uncertainty u^t could be interpreted as the sum of conflicting evidence occurring between 2 or more classes, which could be formulated as

$$u^t = 1 - \sum_{k=1}^{K} b_k^t = \sum_{P, P \in 2^\Theta, 2 \leq |P| \leq K} b_P^t, \tag{3}$$

where we use P as the symbol for any proposition, and it satisfies $0 \leq b_P^t \leq 1$.

4.2. Evidence Fusion

After discussing belief and uncertainty for a single-step observation, we now proceed to explore how to fuse evidence among multi-frame predictions. Two primary fusion strategies are employed in active recognition. The first one is early fusion [29, 55], which recurrently aggregates temporal visual information at the feature level. The second one is late fusion, which involves techniques such as voting or averaging the softmax outputs [16]. We propose fusing per-step evidence within the framework of subjective probabilities [31], which can be considered as a form of late fusion.

The insights of doing late fusion are three-fold. First, the unexpected visual input can occur at any timestep during a recognition episode, regardless of whether it takes place during policy training or testing. Thus, the feature could be contaminated if it is aggregated at an early stage, and difficult to differentiate at which time the recognition uncertainty is introduced. This, in turn, impedes the provision of reasonable rewards for evaluating the policy. Second, late fusion better accommodates the assumption in DST that available evidence from different sources should be independently measured. Third, compared to recurrently fusing temporal features, the late fusion is unaffected by the order of visual observations. Considering a group of visual observations,

the recognition result should remain unchanged by the order in which they are presented.

With no loss of generality, we formulate the evidence combination between any two basic belief functions b_k^t, b_k^j on different observations under the same frame of discernment, *i.e.*, Θ, as:

$$\begin{aligned} b_k = b_k^t \oplus b_k^j &= \frac{1}{\sum_{P^t \cap P^j \neq \emptyset} b_{P^t}^t b_{P^j}^j} \sum_{P^t \cap P^j = k} b_{P^t}^t b_{P^j}^j \\ &= \frac{b_k^t b_k^j + b_k^t u^j + b_k^j u^t}{1 - \sum_{i \neq q} b_i^t b_q^j}, \end{aligned} \tag{4}$$

where P^t and P^j are any two propositions in 2^Θ. Similarly, the uncertainty after the combination is defined as $u = u^t u^j / (1 - \sum_{i \neq q} b_i^t b_q^j)$, where the denominator serves as a normalization factor, representing the total valid mass. It should be noted that the combination reduces conflicting evidence as one single term, *i.e.*, approximate with the uncertainty u^t and u^j as in Equation. 3, to lower the computation complexity. Moreover, this combination guarantees the final prediction to be more strongly influenced by the belief assignment with a lower uncertainty.

Subsequently, given a sequence of T observations collected by the agent, we derive the final belief function by

$$b_k = b_k^1 \oplus b_k^2 \oplus \cdots \oplus b_k^T. \tag{5}$$

The combination inherently adheres to the commutative property, implying that the order of observations does not impact the outcomes. And the final prediction of the category is $\hat{y} = \arg\max_k b_k$.

4.3. Developing Opinions

In this section, we outline how to develop opinions for the visual recognition model using training data. The deep recognition model used in our approach could be generally denoted as a mapping function $f_\theta(\cdot) \rightarrow \mathbb{R}$ with a set of parameters, *i.e.*, θ. The output of the recognition model is obtained directly from the current input as evidence $e_k, 1 \leq k \leq K$ after applying a non-negative function, such as an exponential function or sigmoid. Following evidential deep learning [49], the training is implemented as the evidence acquisition on a Dirichlet prior, whose loss is formulated as

$$\mathcal{L}_{edl}^t = \sum_{i=1}^{K} y_i [\log(\sum_{j=1}^{K} \alpha_j^t) - \log(\alpha_i^t)], \tag{6}$$

where $\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_K]^T$ is the one-hot label, and $\boldsymbol{\alpha}^t = [\alpha_1^t, \ldots, \alpha_i^t, \ldots, \alpha_K^t]$ with $\alpha_i^t = e_i^t + 1$ are parameters of a Dirichlet distribution $Dir(\cdot|\boldsymbol{\alpha})$. Additionally, a Kullback-Leibler loss is incorporated to promote mutual exclusivity among singleton beliefs, defined as

$$\mathcal{L}_{kl}^t = KL(Dir(\cdot|\tilde{\boldsymbol{\alpha}}^t) || Dir(\cdot|\langle 1, \ldots, 1 \rangle)), \tag{7}$$

where $\tilde{\boldsymbol{\alpha}}^t = \mathbf{y} + (1 - \mathbf{y}) \odot \boldsymbol{\alpha}^t$, and $\odot$ is for element-wise multiplication. Combining all terms, the total loss for an observation v^t is

$$\mathcal{L}^t = \mathcal{L}_{edl}^t + \lambda_{kl} \mathcal{L}_{kl}^t, \tag{8}$$

where λ_{kl} is an annealing weight to gradually increase the effect of $\mathcal{L}_{kl}^t$.

Although the combination rule could provide fused predictions, we opt to train our recognition model using single observations only. The rationale behind this choice is that if all observations contain the target to be recognized, adding another loss on fused evidence does not alter the optimality of the loss function. Conversely, if the target is missing in any observations, the training may be adversely impacted.

4.4. Uncertainty-aware Policy Learning

The architecture of proposed agent is demonstrated in Figure. 2. Besides the recognition model to predict the category of the target, the policy module $\pi_\phi(a^t|s^{t-1})$ with parameters ϕ is supposed to control the robot to maximize the accumulated reward $R = \sum_{t=2}^{T} r^t$. The state s^{t-1} describes the aggregated information from observations till the timestep $t-1$. To prevent interference with the recognition, we use a separate recurrent unit $g(\cdot)$ for fusing temporal knowledge. The state is then expressed as $s^t = g(v^1, \ldots, v^t, q_x^t)$, where q_x^t denotes the binary mask of the target x at timestep t. To ensure clarity in our study and to eliminate uncertainties introduced by other modules, the mask q_x^t is derived from the ground-truth bounding box of the target, consistent with previous embodied perception works [40, 53, 57]. In practical applications, the agent can utilize off-the-shelf class-agnostic visual trackers [52] to obtain q_x^t.

We design a novel reward to stabilize policy learning for ambiguous visual observations. According to our prediction result as in Equation 2, the reward r^t should reflect the evidence of the correct class. To this end, the proposed reward is straightforwardly defined as $r^t = b_y^t$. Note that we use b_y^t to represent the belief for the correct class, which can also be interpreted as the normalized evidence. The reward r^t varies between 0 to 1 depending on how much the recognition model collects the evidence for the target class. Since our output includes an uncertainty term, the belief for all known classes may be extremely low when encountering an unexpected input, suggesting that the action leading to the observation should not be rewarded.

As discussed in [19, 55], joint training of the recognition module and the policy can lead to a collapsed outcome. This is because an insufficiently-trained recognition model may not provide accurate rewards during the early training stages. Therefore, we perform a staged training: first, we train the recognition model using a heuristic policy, and then we train only the policy part using PPO [48], while keeping the recognition model fixed.

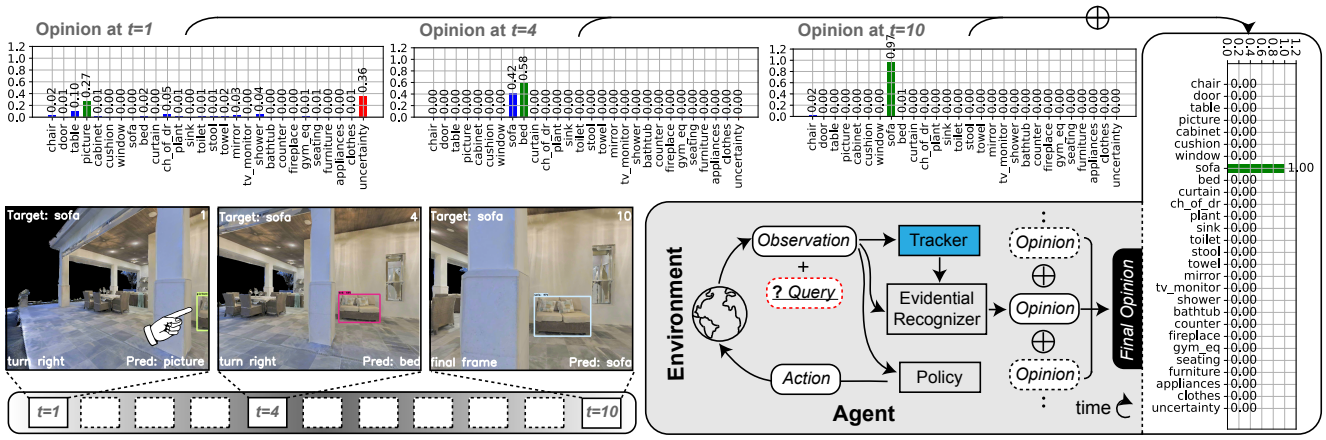

Figure 2. An illustration of a recognition episode and the proposed agent's architecture in the grey box. We select three frames ($t = 1, 4, 10$) along with their estimated opinions. The bars for top prediction and uncertainty are colored green and red, respectively. Note that uncertainty arises when the target is partially out of view at the first step. Despite this, the final result is accurate due to the fusion of evidence.

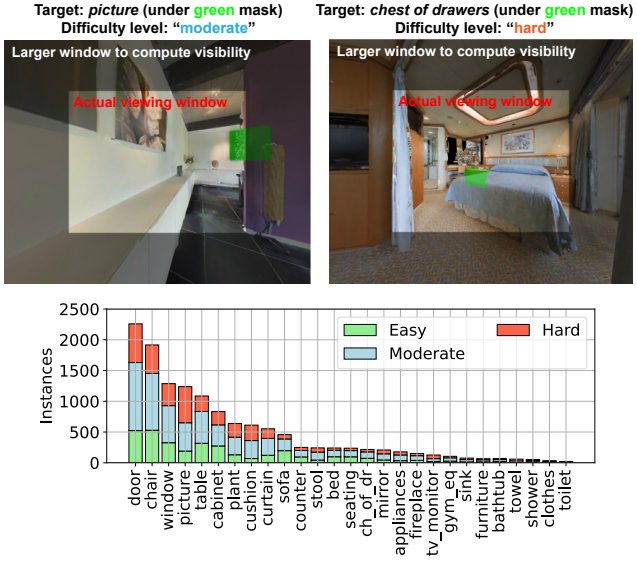

Figure 3. We present two examples of test data from the proposed dataset in the upper sub-figure. The target object is covered by a green mask, allowing for box, point or mask queries during testing. The targets are respectively occluded by the wall and the bed, thus visibility is calculated as the ratio of observed pixels to total pixels belonging to the target. In the lower sub-figure, the distribution of difficulty levels across all categories is depicted.

5. Testing Dataset for Active Recognition

Compared to passive recognition, active recognition is motivated to avoid undesired viewing conditions, including but not limited to heavy occlusions and the out-of-view, where adjusting viewpoints is inevitable. We build a testing dataset designated for assessing the performance of active recognition under different recognition challenges, which is collected from the current simulator MP3D [11, 47, 51].

The testing dataset from MP3D [11] contains 90 building-scale indoor spaces covering diverse scene categories. For each test instance, the agent's starting position and the query are randomly specified, ensuring that the relative distance between them falls within the range of 3 to 6 meters. Since the semantic annotation in MP3D [11] could be noisy, we manually remove erroneous or low-quality queries, resulting in a total of 13200 testing instances. The training of the agent is conducted using the non-overlapping HM3D dataset [43], which contains 145 semantically-annotated indoor scenes, along with randomly generated initial positions and queries.

Object categories. We select 27 object categories from over 40 labeled classes that exists in both datasets [11, 43]. A category is selected only if it does not contain high ambiguities (objects, misc, *etc.*) or belongs to building components (wall, ceiling, stairs, *etc.*).

Recognition difficulty. We introduce an estimate of recognition difficulty for each testing instance. Unlike [55], we evaluate the difficulty level from three perspectives: *visibility*, *relative distance*, and *observed pixels*. *Visibility* is determined by the ratio computed by dividing the unoccluded segmentation mask of the target by its observed mask. To consider partially out-of-view conditions, the unoccluded mask is captured within an enlarged viewing window. Two examples of test instances are depicted at the top of Figure 3, where the left target object is not entirely within the agent's viewing window, and the right target is additionally occluded by a bed. The second aspect, *relative distance*, measures the distance between the agent and the target object. This is determined by subtracting the normalized relative distance (originally ranging from 3 to 6 meters) from 1. *Observed pixels* account for the visible pixels of the target object and are normalized with a cap at 102400 pixels. The final difficulty score combines these three factors using weights of 0.2, 0.2, and 0.6, respectively. We emphasize *observed pixels* be-

cause the other two aspects might not sufficiently represent the recognition challenges posed by tiny objects.

The difficulty level is assigned as follows: "hard" if the score is less than 0.33, "moderate" if it falls between 0.33 and 0.66, and "easy" for all other cases. The difficulty level for each category is depicted in the lower part of Figure. 3. Since we do not re-balance or re-sample rare classes, the category distribution follows a long-tail pattern, reflecting their true occurrences in testing scenes. The dataset will be made publicly available to facilitate reproducible research comparisons. More detailed dataset generation and statistics can be found in the supplementary.

6. Experiments

To validate the effectiveness of our method, we first compare the proposed method with other approaches using our new dataset, demonstrating the benefits of uncertainty-aware policy learning and evidence fusion. Next, we examine the behavior of provided uncertainties across various dimensions, such as categories, steps, and recognition difficulties. More experiments, including ablation studies, are conducted in the final part to further evaluate our method.

6.1. Implementations

The recognition model, based on Faster R-CNN [45], is modified by replacing the region proposal network with the query box at the current timestep. The backbone is ResNet-50 [27] pretrained on the ImageNet [46]. We use the ROI features with a C4 head [45] and also fix the first three residual blocks during training as in [55]. The training of the recognition model incorporates a heuristic fixation coupled with a shortest-path policy, which attempts to approach the target while positioning it at the center of the view. Furthermore, the target is directly provided by the ground-truth bounding box at all timesteps.

The policy part is trained with the recognition model held fixed. The policy network comprises an independent visual encoder, a single-layer Gated Recurrent Unit (GRU) to integrate temporal knowledge, and two single linear-layer to perform as actor and critic on the GRU's output. For detailed architecture and other training hyper-parameters, please refer to our supplementary.

6.2. Baselines

The compared baselines serve two purposes, *i.e.*, to demonstrate the effectiveness of incorporating intelligent strategies into the recognition process, and to highlight the advantages of modeling uncertainties in active recognition. Heuristic recognition policy baselines, including `Single-View`, `Random`, and `Fixation`, employ a similar network architecture, but with alterations to the policy part.

`Single-View`: It imitates conventional passive recognition models that could not intelligently perceiving the target. It uses the initial observation for prediction.

`Random`: The policy randomly selects an action at each step. The number of movements remains the same as ours.

`Fixation`: This policy aims to center the target within the view, potentially reducing undesired out-of-view conditions.

`Amodal-Rec`: Furthermore, we re-implement the embodied amodel recognition agent proposed in [55] which uses an additional convolutional GRU to recurrently aggregate visual features and generate predictions using softmax probabilities. To ensure a fair comparison, we replace the training method in `Amodal-Rec` [55] from REINFORCE [50] to PPO [48] and also the perception part from Mask R-CNN [28] to Faster R-CNN [45], aligning it with our approach.

6.3. Performance Evaluation

In this section, we present a quantitative comparison across active recognition models under various evaluation metrics. In line with established protocols [29, 55], we set a limit of $T = 10$ total allowed steps and subsequently report the performance at the final step. We provide the success rate across varying levels of recognition difficulty. Furthermore, we present a comparative analysis of the success rate from the initial to the final step to illustrate the advantages gained by integrating various movement policies.

In Table. 1, the proposed method achieves improvement over heuristic polices on testing instances with different difficulty levels. For example, the success rate of proposed method is 2.6% higher than `Fixation` policy, while 9.0% higher than `Random` policy. This is primarily because `Random` policy is prone to losing sight of the target easily, while `Fixation` policy does not contribute significant information once the target appears at the center of the view.

Another key observation relates to the results associated with "hard" recognition instances. The increase in success rate for the proposed method on "hard" testing instances ($+8.3\%$) is significantly greater than that on "easy" ($+2.0\%$) and "moderate" ($+2.6\%$) instances. This improvement likely stems from the fact that "hard" instances often involve substantial occlusions and greater distances, conditions that require the agent to adopt more sophisticated viewing strategies in order to acquire clear and distinct observations. We provide a detailed analysis of visibility and distance variations at each step in the supplementary materials.

Furthermore, the aggregation of temporal information at the feature level encounters challenges when deployed in open-world environments. This finding stems from a comparison of the results derived from `Amodal-Rec` [55] with a trained policy versus the corresponding `Single-View` approach. Step-by-step results for `Amodal-Rec` [55] are depicted in Figure 4, revealing a steady trend in recognition as more steps are taken. We postulate two potential reasons for this phenomenon. Firstly, `Amodal-Rec` [55] collates whole-image features across varying frames and sub-

Table 1. Recognition success rates and improvements on the proposed dataset. Success rates are measured according to the final predictions, while the changes in success rates highlight the improvements achieved through movements, thus demonstrating the effectiveness of various policies. `Random` and `Fixation` are two heuristic policies that can be integrated with different recognition agents.

Method	Easy		Moderate		Hard		All		Change in success rate ($t = 1$ to 10)			
	top-1	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	**Easy**	**Moderate**	**Hard**	**All**
`Amodal-Rec + Single-View`	65.1	84.3	55.3	77.1	43.3	66.5	57.4	78.0	-	-	-	-
`Amodal-Rec + Random`	64.7	83.2	55.2	76.8	42.4	66.3	57.0	77.6	-0.4	-0.1	-0.9	-0.4
`Amodal-Rec + Fixation`	64.8	83.3	55.1	76.7	44.4	67.4	57.1	77.6	-0.3	-0.2	+1.1	-0.3
`Amodal-Rec`	65.0	83.5	55.1	76.6	42.9	66.5	57.2	77.7	-0.1	-0.2	-0.4	-0.2
Ours + `Single-View`	67.9	86.2	57.1	77.8	49.7	70.8	60.9	80.7	-	-	-	-
Ours + `Random`	62.0	86.0	49.0	74.9	39.2	67.6	53.4	78.8	-5.9	-8.1	-10.5	-7.5
Ours + `Fixation`	66.2	88.0	58.0	79.0	56.1	78.3	61.8	83.7	-1.7	+0.9	+6.4	+0.9
Ours	**69.9**	**88.3**	**59.7**	**80.4**	**58.0**	**80.2**	**64.4**	**84.3**	**+2.0**	**+2.6**	**+8.3**	**+3.5**

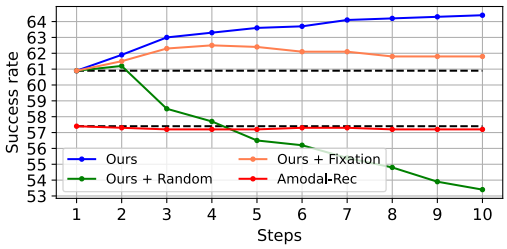

Figure 4. Performance between different agents over steps.

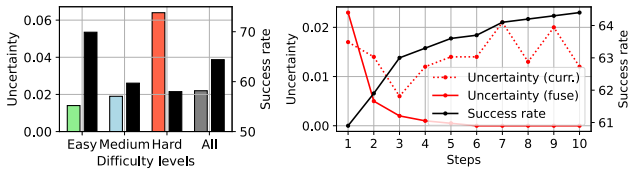

Figure 5. The left illustrates the variation in average uncertainties across different levels of difficulty, with each pair of bars representing the uncertainty and the corresponding success rate. The right section examines the changes in uncertainty.

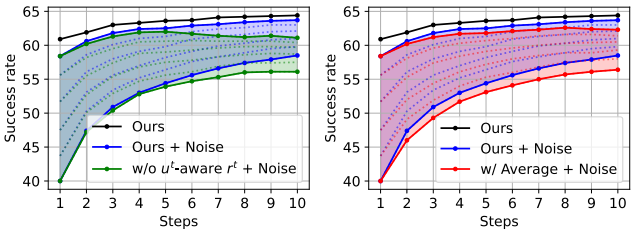

Figure 6. Step-by-step success rates of different variants of the proposed method under varying levels of feature noise. Six noise levels, modeled as Gaussian with $\mu = 0$ and $\sigma = \{2, 3, 4, 5, 6, 7\}$, are added to the features produced by our fixed backbone.

sequently leverages the initial query location to pool regional information from the aggregated feature for classification. As the agent alters its viewpoints, the regional feature may lose its discriminative power as the same queried region incorporates irrelevant information in subsequent frames. Secondly, our dataset presents greater challenges compared to the one used in `Amodal-Rec` [55], primarily due to increased freedom of movement, *i.e.*, the ability to look up and down. Consequently, the likelihood of the same queried region pertaining to the same object between consecutive frames diminishes. These observations underscore the jeopardy of overlooking the inherent challenges posed by active recognition in open-world environments.

The success rate over steps is depicted in Figure 4. Overall, the proposed method shows an improvement in performance as more observations are taken, eventually reaching a saturation point. Interestingly, there is a noticeable performance surge at $t = 2$ for our method with `Random`. This can be attributed to the high probability of the target existing within the observation after a single movement. In essence, the proposed evidence fusion strategy can enhance recognition if the target is observable. Moreover, significant improvements can be achieved through the implementation of our uncertainty-aware policy learning approach.

6.4. Uncertainty in Active Recognition

We delve further into the behavior of uncertainties in our approach. The uncertainty and corresponding final success rate across different difficulty levels are jointly depicted on the left of Figure. 5. For clarity, the uncertainty depicted in the left sub-figure represents the average of the uncertainties at each step, aggregated across all test instances. During testing, as the agent actively acquiring discriminative views for recognition, the uncertainty generally remains at a low level. Furthermore, it can be observed that the uncertainty increases with higher recognition challenges, indicating its potential ability to discern challenging recognition queries.

The trends of step-by-step fused uncertainties and success rates are reported on the right side of Figure. 5. The fused uncertainty decreases and recognition improves as more observations are taken, indicating that the proposed agent effectively gathers evidence to support prediction-making. Additionally, we include the uncertainty before fusion for reference (represented by the dotted red curve), which is independently predicted based on the current view.

Table 2. Ablation studies about different evidence fusion strategies and the integration of uncertainty-aware reward.

Method	Easy	Moderate	Hard	All	
	top-1	*top-1*	*top-1*	*top-1*	*top-3*
Ours	**69.9**	**59.7**	**58.0**	**64.4**	**84.3**
w/ *Max-prediction*	69.3	56.8	56.4	62.9	80.3
w/ *Last-step*	66.2	55.2	55.7	60.7	79.5
w/ *Average*	68.5	57.8	57.5	63.0	82.6
w/ *Vote*	68.6	57.8	56.3	62.8	83.1
w/o u^t-aware r^t	67.5	57.8	53.4	61.6	81.3

6.5. Performance under Feature Disturbance

In this section, we examine the impact of our uncertainty-aware reward strategy during training, and our evidence fusion method, on performance in varied environments. Due to it is impractical and unpredictable to manipulate simulation environments to control recognition challenges, we opted to add Gaussian noise into the visual features obtained by the pre-trained ResNet-50 backbone [27] to simulate unexpected recognition scenarios. These perturbed features are subsequently input into the classification component for per-step prediction, culminating in a fused final output.

We specifically implemented six levels of Gaussian noise, characterized by a mean ($\mu = 0$) and standard deviations ($\sigma = \{2, 3, 4, 5, 6, 7\}$). A higher σ value signifies more intense feature perturbations, presenting greater recognition challenges. Our comparative analysis included two variants of our method, each utilizing the same model architecture. The step-by-step success rates are depicted in Figure 6.

Initially, we evaluated the performance of our method in the absence of uncertainty-aware rewards r^t during policy learning. In this configuration, a binary reward was assigned to the policy based on its current prediction. As illustrated in the left of Figure 6, our approach demonstrated enhanced robustness against feature noise compared to the model trained without uncertainty-aware rewards. We argue that a binary reward system fails to adequately represent the value of actions during training, particularly under conditions of high visual uncertainty, leading to a less effective policy.

Additionally, we substituted our proposed evidence fusion method with an *Average* method, which calculates the mean of estimated beliefs across all frames before making a prediction. This result is showcased in the right of Figure 6. The *Average*, not accounting for uncertainties in each frame, results in an integrated feature that lacks discriminative power, especially in scenarios with high feature noise.

6.6. Ablation Studies

In this study, we explore the impact of different factors such as evidence fusion methods and training rewards on our performance. To eliminate any unrelated interference, we employ the same visual recognition and policy model for all compared methods discussed in this section.

We first evaluate the proposed evidence fusion method against four alternative strategies: *Max-prediction*, *Last-step*, *Average*, and *Vote*. The *Max-prediction* strategy selects the prediction with the highest estimated belief from all steps as the final output, indicating the highest model confidence at that particular step. The *Last-step* strategy solely considers the final single-frame estimation as the outcome. Lastly, the *Vote* strategy implements a voting mechanism among all frame predictions to determine the outcome. Table 2 presents the success rates for various fusion strategies. Our method surpasses these strategies across all levels of recognition difficulty. This superiority primarily stems from the fact that the four alternative strategies fail to account for potential uncertainties arising at each step of embodied recognition, such as occlusions. As detailed in Equation 4, our approach accumulates frame-wise evidence while factoring in estimated uncertainties; thereby, estimates with higher uncertainty exert less influence on the final prediction.

Furthermore, the comparison with agent trained without uncertainty-aware reward emphasizes the efficacy of the proposed reward during training, especially for "hard" testing instances. Essentially, this validates the importance of managing challenges inherent in open environments.

6.7. Limitations and Future Works

In our experiments, recognition performance is evaluated with a fixed number of total steps, yet an ideal agent would be able to determine when to cease taking more movements. At the same time, as both accuracy and the number of observations vary, an effective evaluation metric is further required.

7. Conclusions

In this paper, we examine the challenges faced when deploying active recognition agents in open-world environments, specifically, how to avoid negative impacts from unexpected inputs in class prediction and policy learning. These challenges are inherent in active recognition due to the unpredictable and open nature of the environments being explored. To address this, we propose to model the recognition as a sequential evidence-collecting process, leading to an uncertainty-aware agent. Observations from unknown classes or highly ambiguous views can be rejected, fostering more stable and effective policy learning. A reduced hypothesis space is introduced for evidence fusion, generating the final opinion in accordance with evidence combination theory. A new dataset annotated with recognition difficulties is introduced to evaluate different agents. Experiments on the dataset, along with the uncertainty analyses and ablation studies, confirm the effectiveness of our proposed method.

Acknowledgments

This work was supported in part by National Science Foundation grant IIS-2007613.

References

[1] John Aloimonos. Purposive and qualitative active vision. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, pages 346–360. IEEE, 1990. 1, 2

[2] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4): 333–356, 1988. 1

[3] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. 2

[4] Phil Ammirato, Patrick Poirson, Eunbyung Park, Jana Košecká, and Alexander C Berg. A dataset for developing and benchmarking active vision. In *International Conference on Robotics and Automation*, 2017. 1, 2

[5] Alexander Andreopoulos and John K Tsotsos. A computational learning theory of active object recognition under uncertainty. *International Journal of Computer Vision*, 2013. 1

[6] Nikolay Atanasov, Bharath Sankaran, Jerome Le Ny, George J Pappas, and Kostas Daniilidis. Nonmyopic view planning for active object classification and pose estimation. *IEEE Transactions on Robotics*, 2014.

[7] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42(2):177–196, 2018. 1

[8] Dana H Ballard. Animate vision. *Artificial intelligence*, 1991. 2

[9] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021. 2

[10] Jeffrey A Barnett. Computational methods for a mathematical theory of evidence. *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 197–216, 2008. 3

[11] 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. *arXiv preprint arXiv:1709.06158*, 2017. 5

[12] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 2

[13] Ricson Cheng, Ziyan Wang, and Katerina Fragkiadaki. Geometry-aware recurrent neural networks for active visual recognition. *arXiv preprint arXiv:1811.01292*, 2018. 1, 2

[14] Charles Corbière, Marc Lafon, Nicolas Thome, Matthieu Cord, and Patrick Pérez. Beyond first-order uncertainty estimation with evidential models for open-world recognition. In *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, 2021. 2

[15] Joachim Denzler and Christopher M Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on pattern analysis and machine intelligence*, 24(2):145–157, 2002. 1

[16] Wenhao Ding, Nathalie Majcherczyk, Mohit Deshpande, Xuewei Qi, Ding Zhao, Rajasimman Madhivanan, and Arnie Sen. Learning to view: Decision transformers for active object detection. *arXiv preprint arXiv:2301.09544*, 2023. 1, 2, 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[18] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR, 2020. 2

[19] Lei Fan and Ying Wu. Avoiding lingering in learning active recognition by adversarial disturbance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4612–4621, 2023. 4

[20] Lei Fan, Peixi Xiong, Wei Wei, and Ying Wu. Flar: A unified prototype framework for few-sample lifelong active recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15394–15403, 2021. 1, 2

[21] Lei Fan, Bo Liu, Haoxiang Li, Ying Wu, and Gang Hua. Flexible visual recognition by evidential modeling of confusion and ignorance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1338–1347, 2023. 2

[22] Zhaoyuan Fang, Ayush Jain, Gabriel Sarch, Adam W Harley, and Katerina Fragkiadaki. Move to see better: Self-improving embodied object detection. *arXiv preprint arXiv:2012.00057*, 2020. 2

[23] Samir Yitzhak Gadre, Kiana Ehsani, Shuran Song, and Roozbeh Mottaghi. Continuous scene representations for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14849–14859, 2022. 2

[24] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015. 2

[25] Dimitrios Gallos and Frank Ferrie. Active vision in the era of convolutional neural networks. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 81–88. IEEE, 2019. 1, 2

[26] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

[27] 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. 1, 6, 8

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[29] Dinesh Jayaraman and Kristen Grauman. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. In *European Conference on Computer Vision*, 2016. 1, 2, 3, 6

[30] Dinesh Jayaraman and Kristen Grauman. End-to-end policy learning for active visual categorization. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1601–1614, 2018. 2

[31] Audun Jøsang. *Subjective logic*. Springer, 2016. 2, 3

[32] S Kasaei, Juil Sock, Luis Seabra Lopes, Ana Maria Tomé, and Tae-Kyun Kim. Perceiving, learning, and recognizing 3d objects: An approach to cognitive service robots. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1

[33] Klemen Kotar and Roozbeh Mottaghi. Interactron: Embodied adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14860–14869, 2022. 2

[34] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020. 2

[35] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020. 2

[36] Huaping Liu, Yupei Wu, and Fuchun Sun. Extreme trust region policy optimization for active object recognition. *IEEE transactions on neural networks and learning systems*, 29(6): 2253–2258, 2018. 2

[37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 1

[38] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016. 2

[39] David Nilsson, Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Embodied visual active learning for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2373–2383, 2021. 2

[40] Anwesan Pal, Yiding Qiu, and Henrik Christensen. Learning hierarchical relationships for object-goal navigation. In *Conference on Robot Learning*, pages 517–528. PMLR, 2021. 4

[41] Santhosh K Ramakrishnan and Kristen Grauman. Sidekick policy learning for active visual exploration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 413–430, 2018. 1

[42] Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. Emergence of exploratory look-around behaviors through active observation completion. *Science Robotics*, 2019. 2

[43] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 5

[44] Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *International Journal of Computer Vision*, 129(5):1616–1649, 2021. 2

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 6

[47] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 2, 5

[48] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 4, 6

[49] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 2, 3, 4

[50] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*. MIT press Cambridge, 1998. 6

[51] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021. 2, 5

[52] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 4

[53] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[54] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2

[55] Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2040–2050, 2019. 1, 2, 3, 4, 5, 6, 7

[56] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici.

Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 1

[57] Qianfan Zhao, Lu Zhang, Bin He, Hong Qiao, and Zhiyong Liu. Zero-shot object goal visual navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2025–2031. IEEE, 2023. 4