Treasure Hunting: Embodied Contrastive Learning-Enhanced Coarse-to-Fine Object Seeking With Explorer and Discriminator Cooperation

Bolei Chen, Jiaxu Kang, Haonan Yang, Ping Zhong, *Member, IEEE*, Rui Fan, *Senior Member, IEEE*, and Jianxin Wang, *Senior Member, IEEE*

***Abstract*—Object navigation (ObjcetNav), which enables an agent to seek any instance of an object category, has shown great advances. However, current agents are built upon occlusion-prone visual observations or compressed 2-D maps, which hinder their embodied perception of 3-D scene geometry. Furthermore, existing methods usually decouple ObjectNav into the exploration and exploitation subtasks, easily leading to ambiguous object localization and blind exploration. To address these issues, we first propose an embodied contrastive learning (ECL) method with geometric consistency (GC) and behavioral awareness (BA), which motivates agents to encode 3-D scene layouts and semantic cues actively. The BA is modeled by predicting navigational actions based on multiframe visual images, as behaviors causing differences between adjacent visual sensations are crucial for learning correlations among continuous visions. The GC is modeled by aligning the behavior-aware visual stimulus with 3-D semantic shapes through unsupervised contrastive learning. Then, based on the above ECL pretraining, a coarse-to-fine ObjectNav policy with explorer and discriminator cooperation is proposed, inspired by the treasure-hunting mindset. Concretely, the explorer is designed to adaptively switch the action spaces, thereby switching the global and local exploration thoughts according to the accumulated scene priors. The discriminator is designed to discriminate the target's authenticity using behavior-aware visual features and geometric invariance priors, which permits mimicking the human behavior of "approaching to confirm" when distinguishing objects from a distance. As expected, our ECL method performs well on object detection (ObjDet) and instance segmentation (InstSeg) tasks. Our ECL-enhanced ObjectNav strategy outperforms state-of-the-art (SOTA) methods on Matterport3D (MP3D), Gibson, and HM3D datasets.**

***Index Terms*—Behavioral awareness (BA), embodied contrastive learning (ECL), geometric consistency (GC), object localization, object navigation (ObjcetNav), scene exploration.**

Received 19 November 2024; revised 12 April 2025 and 1 July 2025; accepted 7 October 2025. Date of publication 20 November 2025; date of current version 3 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62272489, Grant 62332020, and Grant 62350004; in part by the Natural Resources Science and Technology Plan Project of Hunan Province under Grant 2021-17; and in part by the Open Competition Project of Xiangjiang Laboratory under Grant 23XJ01011. *(Corresponding authors: Jianxin Wang; Ping Zhong.)*

Bolei Chen, Jiaxu Kang, Haonan Yang, Ping Zhong, and Jianxin Wang are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: ping.zhong@csu.edu.cn; jxwang@mail.csu.edu.cn).

Rui Fan is with the College of Electronics and Information Engineering, Tongji University, Shanghai 200092, China (e-mail: ranger_fan@outlook.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2025.3619968, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2025.3619968

I. Introduction

OBJECT navigation (ObjectNav) task [1], [2] requires an agent to navigate through a previously unknown 3-D scene to find an object instance, according to a semantic label. Existing work has made great advances in visual representations [3], [4], [5], [6], data augmentation techniques [7], [8], and auxiliary tasks [9], [10] for pretraining. Their core ideas are fully exploiting scene layouts and semantic contexts to enhance agents' object localization or scene exploration capabilities. Some methods [3], [4], [11], [12] speculate on correlations among historical visual features for ObjectNav decision-making by emphasizing spatiotemporal awareness of visual observations. Although promising progress has been made, domestic scenes are characterized by substantial occlusion, which poses challenges for agents to accurately localize object goals and efficiently explore scenarios. Moreover, agents typically establish high-level awareness of objects by moving around and perceiving them from different angles and distances. For instance, learning about basic physical concepts for object localization, such as large and long, requires moving beyond image-based observations.

Research in behavioral psychology [13], [14] has shown that many animals maintain spatial representations of their environments while navigating. Inspired by this, some other methods attempt to develop topological scene representations (TSRs) [5], [15], [16], [17], [18], [19] or 2-D contextual semantic maps [1], [20], [21], [22], [23] based on visual images to balance exploration and exploitation better, as shown in Fig. 1(a). Nodes in TSRs typically consist of abstract visual or object features [18], [19]. Edges in TSRs usually involve discrete semantic or geometric relationships (e.g., the pillows are in the bed and the mouse is used to operate the computer) [15], [16]. However, the abundant geometric and semantic relations among objects should be a large relational space and thus difficult to model with discrete TSRs exhaustively. The 2-D contextual semantic maps somewhat reconstruct the layouts and semantic patterns of the scenarios and can provide agents with compressed materials to formulate the continuous relational space [24]. However, RGB image-based semantic segmentation errors may lead to low-quality and ambiguous 2-D semantic maps, which severely impair the ObjectNav performance (see Appendix A, Supplementary Material for more details).

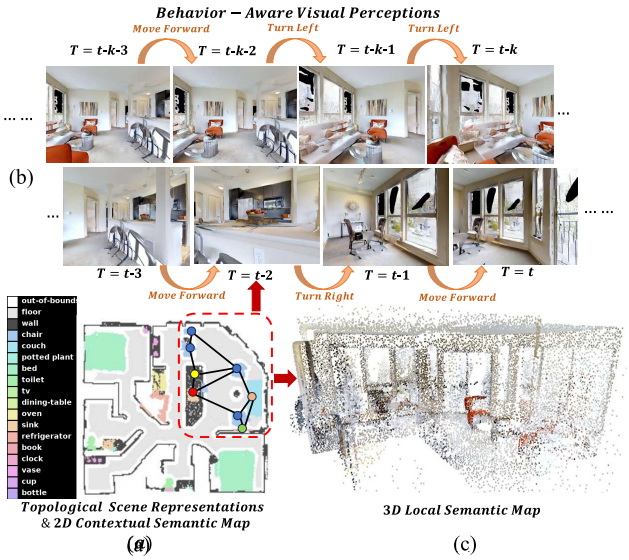

Fig. 1. (a) Illustration of a 2-D contextual semantic map and a local TSR. (b) Illustration of behavior-aware visual perceptions. The changes in visual sensations are caused by navigational behaviors. (c) Illustration of 3-D local semantic maps used for geometric-aware contrastive representation learning, corresponding to the local region in (a).

Beyond the abovementioned efforts in scene representation, existing works [20], [22], [23], [25], [26] have also proposed impactful solutions by decoupling ObjectNav into scene exploration and object recognition subtasks in the exploration–exploitation framework. However, these methods still suffer from two thorny challenges: 1) inefficient scene exploration and 2) blind object recognition. Exploration [27], [28] is essential for agents to feel and recognize their surroundings. Cutting-edge methods [1], [22], [23], [26] are devoted to mining spatial and semantic co-occurrence relations among objects in top-down semantic maps to provide either explicit local frontier guidance (Guid-F) [22] or ambiguous global orientation guidance [26] for scene exploration. However, their exploration subpolicies are unilaterally designed since they empower the agent with a single inefficient exploration capacity. They are either not meticulous enough, causing the agent to wander globally in areas already explored, or too meticulous, causing the agent to fall into localized traps.

In addition, it is extremely challenging to enable the agent to correctly identify the target object from a considerable distance. Existing methods [20], [22], [23], [25], [29], [30] cleverly project instance segmentation (InstSeg) in images onto 2-D semantic maps to localize object targets. However, once the target semantic appears in the incrementally updated semantic map, existing methods [20], [22], [23], [29] directly identify the corresponding region as the object target and recklessly drive the agent navigating to it. Since scene-specific InstSeg is a challenging problem in itself, erroneous segmentation may lead to ambiguities in the semantic map, which further leads to ObjectNav failures. These methods blindly identify the original semantic segmentation as the target object without additional thinking and ignore the discrimination process across the temporal dimension.

To alleviate the above problems, we first propose an embodied contrastive learning (ECL) method with geometric consistency (GC) and behavioral awareness (BA) to motivate agents to actively explore 3-D scene layouts and encode semantic cues. Instead of modeling the correlations among visual features from a spatiotemporal aware perspective, we advocate inferring the relevance among visual features at the root by predicting intermediate actions from consecutive visual frames. We believe that the behaviors that lead to differences between two adjacent observations are crucial for learning the relations between two visions, as shown in Fig. 1(b). Our BA modeling is more concise than predicting visual features from action sequences since the observation space consisting of high-dimensional sensor inputs tends to be large and variable, while the action space is small, discrete, and relatively fixed.

When searching for a specific object, humans usually take sequential actions to actively transform their field of view (FoV), locating the target and exploring the scene by aligning their visual senses with 3-D space. Inspired by this, we empower agents to reconstruct local 3-D spatial structures and semantic patterns during navigation, as shown in Fig. 1(c). The 3-D local scene priors allow agents to learn rich scene representations by immersively aligning the visual features encoded by the behavior-aware visual encoder with the 3-D scene features encoded by the 3-D point cloud (PCL) encoder. As a result, the 2-D scene understanding is enhanced by introducing geometric and view-invariant priors into the behavior-aware 2-D visual features. In a nutshell, the GC is modeled as aligning behavior-aware visual perceptions with 3-D semantic shapes by employing unsupervised contrastive learning.

Notably, to adequately mimic the situational interactions of humans with 3-D scenes, the above BA and GC-based contrastive representation learning is performed in an embodied manner. In particular, we propose a curiosity and action-aware exploration policy named E^2-CL, which continually motivates agents to adopt diverse actions to discover novel visual perceptions. On the one hand, the semantic-rich visual stimulus facilitates agents to carry out more comprehensive and robust 2-D–3-D scene representation learning. On the other hand, the diverse action–vision data pairs collected online provide rich learning materials and feature bases for BA modeling. With the collection of novel and complicated action–vision pairs, the visual frame-based action prediction will be more challenging. Therefore, the BA modeling will be gradually enhanced in this adversarial learning process.

Inspired by the classic task decoupling [7], [22] and treasure-hunting thinking, we further propose a coarse-to-fine ObjectNav policy with explorer and discriminator cooperation to mitigate inefficient scene exploration and blind object recognition. Please imagine an explorer and a discriminator together on a treasure hunt in a complex and previously unknown scenario, as shown in Fig. 2(a). One of our core points is that the excellent insight into clues and the cooperative awareness of humans may inspire novel solutions. In addition, we emphasize using a coarse-to-fine thought to tackle the exploration and localization subtasks.

Specifically, a dual-thinking explorer is proposed to change exploratory thinking by changing action spaces based on historical semantic and 3-D geometric clues. Similar to humans

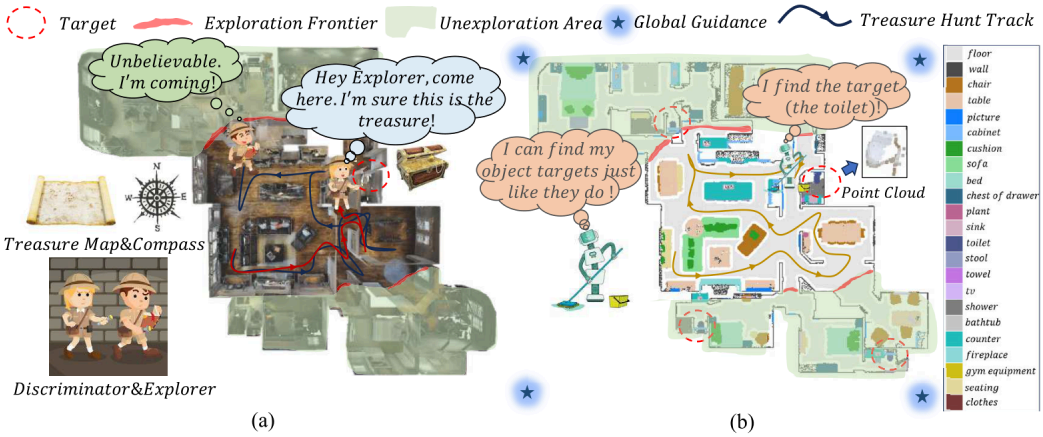

Fig. 2. (a) Brave explorer proactively unravels the mysteries of the scenario and constantly inspects new areas based on the clues provided by the treasure map and the guidance of the compass. The careful discriminator cautiously identifies the authenticity of the treasure, leaving no stone unturned in its possibilities. (b) Inspired by the treasure-hunting mindset, ObjectNav is solved as the tacit collaboration of the explorer and the discriminator.

who explore a scene by aligning their behaviorally aware visual senses with 3-D space, the explorer benefits from BA and GC modeling in ECL pretraining. Our explorer can adaptively switch between fine-grained local exploration and holistic global exploration, just as the explorer decides whether to inspect the local surroundings or use compass guidance to open up new spaces. In other words, the explorer fully explores the frontiers in a local region when there is a high likelihood that a target exists in the region **(fine stage)**. If a local area is thoroughly inspected but still no target is found, the explorer switches to global exploration mode to find the next most promising local area **(coarse stage)**.

Following a novel discrimination–exploitation guideline, a coarse-to-fine discriminator is proposed to identify the object target's authenticity while approaching it. Notably, we utilize the behavior-aware visual features and geometric invariance priors modeled by ECL to learn the discriminatory criterion, instead of using the raw visual images. The reason is that images are inherently viewpoint-specific and scale-agnostic and fail to capture the 3-D objects' physical extent and locations in the scene. Once the object target's semantic emerges in the metric map, the agent is driven toward the corresponding region to verify whether the target semantic is correct **(coarse stage)**. If the 3-D geometric feature-based discrimination fails, the region's semantic is eliminated to prevent misleading the ObjectNav task **(fine stage)**. Overall, our ObjectNav agent follows a novel exploration–discrimination–exploitation guideline and was born from the collaboration of explorer and discriminator.

During the experimental phase, we first validate the superiority of our ECL on generic object detection (ObjDet) and InstSeg tasks. Particularly, the visual encoder pretrained by ECL is further retrained to solve these two tasks. The great performance of our approach on both tasks reflects ECL's expertise in object recognition, which will be further migrated to the ObjectNav task. In addition, the pretrained visual and PCL encoders are integrated into our coarse-to-fine ObjectNav strategy, which is compared with state-of-the-art (SOTA) ObjectNav methods on Matterport3D (MP3D) [31], Gibson [32], and HM3D [33] datasets. Sufficient ablation studies demonstrate the substantial contributions of the individual components in our method. Overall, the contributions of this article are as follows.

1) An embodied contrastive representation learning paradigm with BA and GC is proposed. The BA modeling helps agents take informed navigational actions based on behavior-aware visual perceptions. The GC modeling infuses agents with 3-D geometric invariance priors.
2) A curiosity and action-aware exploration policy is proposed to support embodied ECL. The diverse action–vision pairs collected online provide rich feature bases for the BA and GC modeling. The training of embodied exploration policy and ECL are adversarial but enhance each other.
3) A novel learning system is proposed for ObjectNav. The dual-thinking explorer can explore unknown scenes effectively by adaptively switching exploratory thinking. The prudential discriminator can imitate the human behavior of "approaching to confirm" when distinguishing objects from a distance.
4) Sufficient comparative and ablative studies on ObjDet, InstSeg, and ObjectNav tasks demonstrate the superiority of our ECL, E^2-CL, and coarse-to-fine ObjectNav methods. The experimental code will be open-sourced after the anonymous review.

II. Related Work

A. Scene Representation for Visual Navigation

1) Spatiotemporal Visual Modeling for Visual Navigation: Human beings can naturally navigate in new environments, which requires us to find parallels between the new observations and our past experiences. Inspired by this, visual modeling of spatiotemporal awareness [3], [4], [11], [12], [34], [35], [36], [37] can provide agents with historical contextual information for navigation. Its core concept is to implicitly encode the visual semantic clues, the relative spatial information among objects, and the temporal correlations among multiple visual frames, using recurrent neural networks [3], [4], [11] or Transformers [12], [34], [35], [36], [37]. Some improved methods [11] exploit spatiotemporal attention mechanisms to filter keyframes and intelligently focus on semantic

and spatial cues that are most relevant to the navigation goal. One of our main insights is that variations in visual perceptions are the consequence of active navigational behaviors. Therefore, unlike spatiotemporal visual modeling that extrapolates dynamic correlations back from observations, we believe that the modeling of behavioral patterns can help characterize those correlations at the source.

2) TSRs and Semantic Maps for Visual Navigation: TSRs have continuously shown improvements in various visual semantic tasks. Visual language navigation [38], [39], [40], [41] typically embeds panoramic visual features into the topological graph's nodes and represents the topological graph's edges as combinations of relative positions and orientations between the nodes. Although the topology retains the partial geometry of the scene, the relationships and dynamic transitions between nodes should be encoded with richer geometric and semantic cues. ObjectNav [3], [4], [42] usually injects object features detected from visual perceptions into the topological graph's nodes. The difference is that the topological graph's edges include not only geometric relative positions but also semantic relationships, e.g., a mouse is used to operate a computer. Nevertheless, some studies of continuous scene representations [24], [43] have shown that such discrete relational modeling is prone to inductive bias in spatial and semantic understanding.

Since semantic maps preserve fine-grained scene layouts and semantic patterns, they can alleviate inductive bias by providing agents with richer scene representations. By projecting high-dimensional features encoded from neural networks to a top-down map, existing works [44], [45] try to generate a deep feature map, which is used for reconstructing scenes, predicting semantic maps, and navigating. The semantic map is a type of occupancy map [46], which indicates whether a point is occupied, represents the objects' semantic categories, and provides location information for robotic navigation and exploration [47], [48]. By decoupling ObjectNav tasks into object localization and scene exploration subtasks, a series of 2-D semantic map-based modular methods [22], [23], [49], [50] has been proposed and achieved promising performance. Although 2-D contextual semantic maps somewhat reconstruct the layouts and semantic patterns of the scenarios, they lose the 3-D geometric structure that is critically important for embodied navigation.

Although existing work achieves impressive performance on visual navigation by introducing 3-D semantic maps [26], [51] and bird's-eye views [52], [53], how to fully mine and exploit 3-D geometric features is still a challenging and open topic. To release the agents from the 2-D observation space, this article proposes a novel 2-D–3-D contrastive representation learning method with geometric awareness by aligning rich visual features with 3-D scene priors in an embodied manner.

3) Unimodal and Multimodal Representation Learning for Visual Navigation: Unimodal representation learning has emerged with significant success in visual navigation tasks. OVRL [54] employs the concept of knowledge distillation [55] to learn navigation-friendly visual representations from pure visual images in an offline manner. In contrast, CRL [43] presents an embodied adversarial contrastive learning technique to encourage agents to actively explore novel surroundings for learning robust visual representations online. To alleviate the typically substantial occlusions in visual images, Chen et al. [24] employed offline contrastive learning to extract continuous relations among objects from 2-D semantic maps. In terms of TSRs, a novel scene graph contrastive loss [56] is proposed to encourage representations that encode objects' semantics, relationships, and history. Unlike the above methods, one of our main insights is to merge the merits of different modalities to achieve richer and more robust scene representations.

Multimodal representation learning gains attraction due to its ability to share modality-specific contexts. Humans naturally build spatially meaningful cognitive maps in their brains during navigation. Inspired by this, Ego^2-Map [57] proposes a navigation-specific method for learning visual representations by aligning egocentric views with 2-D semantic maps in a cross-modal manner. Although this idea is commendable, ObjectNav agents are born and work in 3-D scenes. In this work, we experimentally prove that 3-D geometric awareness is crucial for ObjectNav decision-making by comparing our method with Ego^2-Map. Inspired by recent work on 2-D–3-D multimodal representation learning [58], [59], [60], [61], this article proposes a GC-based ECL technique that encourages agents to actively exploit 3-D geometric and semantic priors. Our inspiration comes from Pri3D [60], which introduces a contrastive learning method for multiview RGB frames and 3-D scene scans. Unlike Pri3D, our ECL method can organically fuse the geometric features in 3-D semantic shapes with behavior-aware visual features. The features from both modalities will be utilized to enhance ObjectNav. In addition, our ECL is conducted online in an embodied fashion, with the aim of active learning through interactions with the scenes.

B. Scene Exploration for ObjectNav

Although end-to-end ObjectNav strategies [5], [11], [42], [62], [63] do not emphasize scene exploration explicitly, the navigational actions made in response to RGB visual observations inherently involve the process of gradually exploring and discovering novel visual stimuli [28]. The exploratory skills are more evident in the modular approaches [20], [22], [23], [25], [26]. For example, SemExp [20] pioneers using the entire semantic map as the exploration strategy's action space to regressively predict an interim exploration goal. Recently, Stubborn [25] and 3DAware [26], respectively, propose heuristic and learning-based corner-guided global exploration strategies to address the inefficient sampling caused by SemExp's large action space. Alternatively, PONI [22] and PEANUT [23], respectively, propose frontier potential- and region potential-based exploration policies to provide agents with explicit local guidance. Despite the promising results, these methods empower the agent with a single inefficient exploration capacity, which may lead to ineffective wandering or falling into localized traps. To alleviate this issue, this article proposes a dual-thinking explorer capable of switching exploratory mindsets to achieve effective scene exploration.

C. Object Localization for ObjectNav

End-to-end ObjectNav strategies [5], [11], [42], [62], [63] typically employ ObjDet models such as Msak-RCNN [64]

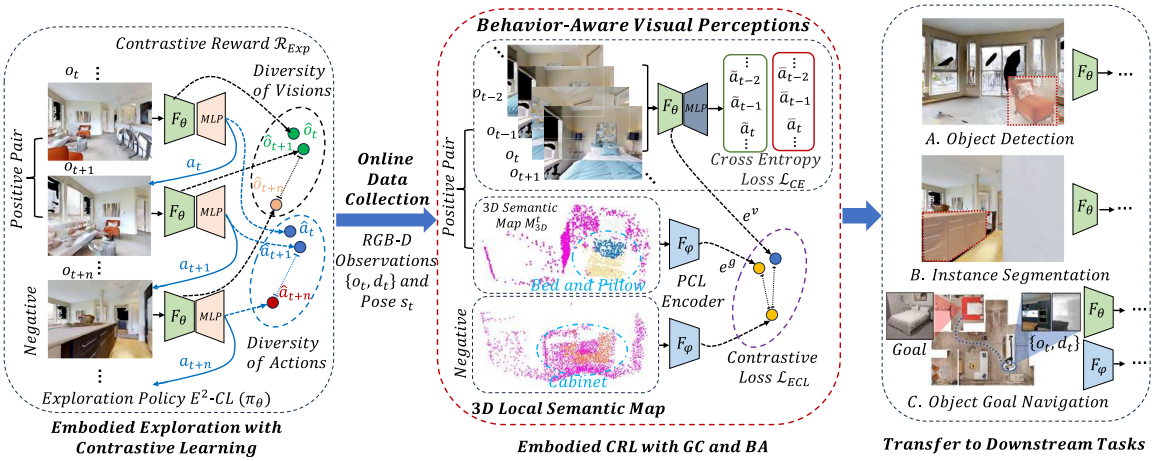

Fig. 3. (Left: Section III-B) Driven by the exploration policy π_θ, our embodied agent actively seeks novel visual observations by adopting diverse actions to maximize the curious contrastive reward $\mathcal{R}_{\text{Exp}}$. (Middle: Section III-C) Our ECL method aligns the behavior-aware visual features with the 3-D geometric features to model GC by minimizing the contrastive loss $\mathcal{L}_{\text{ECL}}$. The BA is modeled by minimizing the cross-entropy loss $\mathcal{L}_{\text{CE}}$ between the predicted actions and the real actions. (Right) Pretrained F_θ and F_φ are transferred to downstream tasks for retraining and task-specific performance evaluation.

to identify objects from raw RGB visual images. However, it is extremely challenging to empower an agent to correctly recognize the target object at a considerable distance, especially when the images are occluded. In contrast, the modular methods [20], [22], [23], [29] localize object targets indirectly by projecting visual InstSeg onto 2-D semantic maps and checking whether the target semantic appears in the metric maps. However, they typically treat the decision-making process as a one-shot task and overlook the judgment process across the temporal dimension. In particular, they ignore the semantic ambiguity caused by semantic segmentation errors and recognize the target region blindly, which leads the agent to find the wrong object. To alleviate this issue, this article proposes a prudential discriminator to imitate the human behavior of "approaching to confirm" when distinguishing objects from a distance, following a novel discrimination–exploitation guideline.

III. Methodology

A. Problem Statement and Overview

1) Problem Statement: In an unknown environment, the ObjectNav task requires the agent to navigate to an instance of the specified target category. As initialization, the agent is located randomly without access to a prebuilt environment map and is given a target category $c_{\text{target}} \in \{1, 2, \ldots, C\}$, where C is the number of possible target categories. For each navigation state s_t, the agent receives noiseless onboard sensor readings, including egocentric RGB-D images $\{o_t, d_t\}$ and a 3-DoF pose $\{x, y, \theta\}$ (2-D position and 1-D orientation) relative to the starting of the episode. Then, the agent predicts its action a_t for movement in a discrete action space, consisting of Move_Forward, Turn_Left, Turn_Right, and Stop. An episode is considered successful if the agent executes Stop within 1.0 m of a target object and the object can be viewed from the agent's position. Each episode has a time limit of 500 steps.

2) Overview: Figs. 3 and 4 provide the overviews of the proposed ECL method and our modular ObjectNav strategy, respectively. A curious contrastive reward $\mathcal{R}_{\text{Exp}}$ with action awareness is designed for the embodied exploration in ECL, which motivates agents to actively explore the scene and consistently gather novel visual images (Section III-B). We believe that navigation behavior is one of the main factors affecting embodied agents' visual perceptions. The BA is modeled by using continuous visual frames to predict the intermediate actions (Section III-C). Meanwhile, a PCL-based semantic mapping method is employed to project the semantically segmented RGB images to a 3-D semantic map based on the depth images, the camera parameters, and the agent's poses. This article focuses on whether 3-D scene priors can enhance visual semantic navigation; thus, a contrastive loss $\mathcal{L}_{\text{ECL}}$ is proposed for GC modeling by aligning behavior-aware visual features with corresponding 3-D scene priors (Section III-C).

In addition, to mitigate inefficient scene exploration, an explorer is designed to adaptively switch the action spaces, thereby switching the global and local exploration thoughts according to the accumulated scene priors (Section III-D). To address blind object recognition, a discriminator is designed to discriminate the target's authenticity using behavior-aware visual features and geometric invariance priors, which permits mimicking the human behavior of "approaching to confirm" when distinguishing objects from a distance (Section III-E). Finally, the implementation details of the ObjectNav policy are presented (Section III-F).

B. Embodied Exploration With Contrastive Learning (E^2-CL)

This work advocates that agents learn to build environmental cognition by continuously interacting with their surroundings as humans do. To ensure adequate interactions with diverse scenarios, a curiosity-driven exploration policy π_θ named E^2-CL is designed to motivate embodied agents to actively explore the scenarios. π_θ drives agents to consistently discover novel visual perceptions and facilitates the learning of semantically enriched scene representations. Specifically, the exploration policy π_θ consists of a visual encoder F_θ and a multilayer perceptron (MLP) projection head used for predicting exploration actions, as shown in Fig. 3 (left). For the RGB

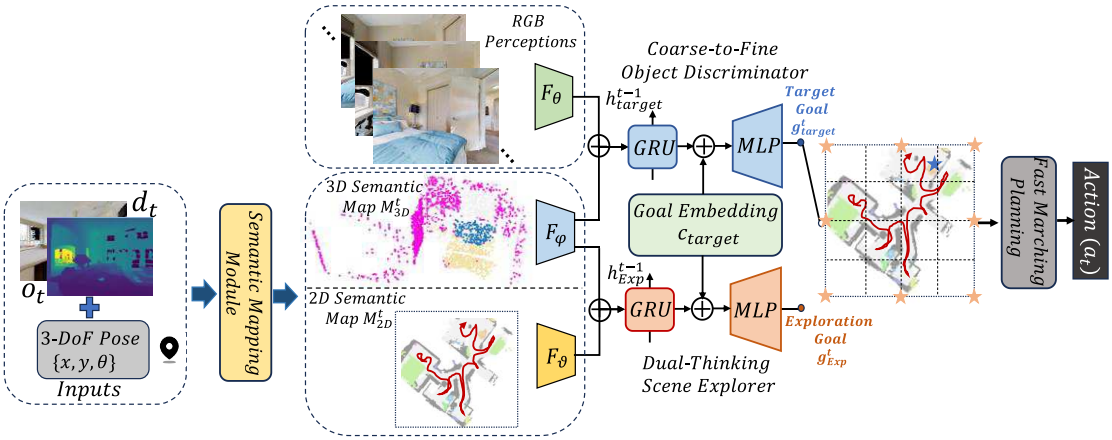

Fig. 4. ObjectNav strategy takes RGB-D images, agent's pose, and object goal category as inputs. A PCL-based semantic mapping module is employed to build a 3-D semantic map M_{3D}^t along with a projected 2-D semantic map M_{2D}^t based on semantically segmented RGB-D images and poses. The pretrained visual encoder F_θ and the PCL encoder F_φ are utilized for object localization. F_φ and another 2-D map encoder F_ϑ are used for scene exploration. A deterministic local planning policy is utilized to drive the agent to the target goal g_{target}^t or exploration goal g_{Exp}^t.

visual observations $O = \{o_k\}_{k=1,2,\ldots,N}$ collected by the agent during exploration, different image augmentation methods are applied to each image to obtain pairs of augmented images $\hat{O} = \{\hat{o}_k^1, \hat{o}_k^2\}_{k=1,2,\ldots,N}$. N denotes the number of collected RGB images. Each augmented image is encoded using F_θ and followed by L2 normalization, which is formulated as $\hat{z}_k^* = \text{Normalize}(F_\theta(\hat{o}_k^*))$. To maximize the diversity of visions, the following reward signal is used to train π_θ by maximizing the similarity between the augmented images in the same pair and minimizing the similarity between the augmented images in different pairs:

$$\mathcal{R}_V = \frac{1}{N}\sum_{k=1}^{N} \log \frac{\exp\left(\text{sim}\left(\hat{z}_k^1, \hat{z}_k^2\right)/\tau\right)}{\sum_{z^- \in \hat{Z}\setminus\{\hat{z}_k^1\}} \exp\left(\text{sim}\left(\hat{z}_k^1, z^-\right)/\tau\right)} \tag{1}$$

where $\hat{Z} = \{\hat{z}_k^1\}_{k=1,2,\ldots,N}$, τ is a softmax temperature scaling parameter, and $\text{sim}(\cdot,\cdot)$ corresponds to the dot product.

In practice, we find that the agent may only slyly perform steering actions in place to greedily maximize the reward described in (1). To alleviate this problem, another reward signal is designed to maximize the diversity of actions

$$\mathcal{R}_{\mathcal{A}} = \frac{1}{N}\sum_{k=1}^{N} \sum_{a^- \in \hat{\mathcal{A}}\setminus\{\hat{a}_k\}} \text{sim}\left(\hat{a}_k, a^-\right) \tag{2}$$

where $\hat{\mathcal{A}} = \{\hat{a}_k\}_{k=1,2,\ldots,N}$, $\hat{a}_k = \text{MLP}(a_k)$, and $\mathcal{A} = \{a_k\}_{k=1,2,\ldots,N}$ denotes the actions taken to collect visual observations. Overall, the curiosity and action awareness-driven reward $\mathcal{R}_{\text{Exp}} = \mathcal{R}_V + \alpha\mathcal{R}_A$ is employed as the total reward signal for our E^2-CL. α is a weight used to balance the two rewards. By doing so, the agent is motivated to discover diverse RGB images by taking diverse actions.

Notably, the exploration policy is trained to maximize the following cumulative reward by utilizing the proximal policy optimization (PPO) [65]:

$$\max_\theta \mathop{\mathbb{E}}_{x\sim\pi_\theta}\left[\sum_{t=0}^{T} \mathcal{R}_{\text{Exp}}(F_\theta, x)\right]. \tag{3}$$

In particular, π_θ is trained by optimizing the objective $L(\theta) = \mathbb{E}[\min(c_t(\theta)A_t, \text{clip}(c_t(\theta), 1-\epsilon, 1+\epsilon)A_t)]$, where the clip ratio $c_t = (\pi_{\theta(a_t|s_t)})/(\pi_{\theta_{\text{old}}}(a_t|s_t))$ and the advange A_t is computed utilizing the value function $V(s_t)$. During exploration, π_θ is optimized by using the collected minibatches of data from PPO (see Appendix C for more details). As shown in Fig. 3, the RGB-D images $\{o_t, d_t\}$ and agent's poses s_t collected by E^2-CL are used for the subsequent online ECL.

C. ECL With GC and BA

1) BA Modeling: When searching for a specific object, humans usually actively transform their FoVs or move forward to localize the object instance. Inspired by this, we propose to model the correlations between behaviors and visions (the long-horizon dynamic transitions between visual frames) by predicting action sequences based on successive multiframe visual perceptions, as shown in Fig. 3. Notably, our exploration strategy E^2-CL tends to adopt diverse actions to discover novel visual stimuli, which provides rich behavior-vision data for BA modeling. To be more specific, a neural network consisting of the visual encoder F_θ and an MLP projection head is utilized to predict l intermediate navigation actions $\{\tilde{a}_i\}_{i=t}^{t+l-1}$ from $l+1$ RGB visual frames $\{o_t\}_{i=t}^{t+l}$. This procedure is accompanied by the embodied exploration so that the agent's real navigation actions $\{\bar{a}_i\}_{i=t}^{t+l-1}$ can be collected as supervisory signals. The cross-entropy loss $\mathcal{L}_{\text{CE}} = -1/l\sum_{i=0}^{l-1}\bar{a}_i\log\tilde{a}_i$ is used to optimize the neural network.

Notably, the agent's action space is small and discrete, while the observation space is relatively large and variable. Therefore, our BA modeling is concise and manageable in complexity compared to the modeling technique of predicting visual observations from navigation actions.

2) GC Modeling: To align the 2-D behavior-aware visual features with the 3-D scene priors, the semantically segmented RGB images are projected into a 3-D semantic map in the form of PCL based on the camera's parameters and the corresponding depth images (Please see Appendix B for more details). At time step t, the past $l+1$ frames of visual observations $\{o_i, d_i\}_{i=t-l}^t$ and poses $\{s_i\}_{i=t-l}^t$ are utilized to construct a 3-D local semantic map $\mathcal{M}_{3\text{D}}^t \leftarrow \{(P_p^t, P_s^t)\} \in \mathbb{R}^{Q^t\times(C+3)}$ for the consistency of 2-D–3-D features. Here, $P_p^t \in \mathbb{R}^3$ and $P_s^t \in \mathbb{R}^C$

denote each point's position and semantic category, respectively. C and Q denote the number of semantic categories and the number of PCL in $\mathcal{M}_{3D}^t$, respectively. In practice, the off-the-shelf models [64], [66] are employed to obtain semantic segmentations from $\{o_i\}_{i=t-l}^t$ and combine the depths $\{d_i\}_{i=t-l}^t$ to generate 3-D semantic PCL. However, each semantic point may probabilistically belong to multiple different semantic categories. Therefore, we suggest employing a max-fusion mechanism [26] to merge the temporal-variant semantic PCL to ensure consistent scene semantics.

Humans often efficiently localize specific object targets in an embodied manner by actively matching their visual senses with 3-D scene structures. Inspired by this, a multimodal contrastive loss $\mathcal{L}_{\text{ECL}}$ is proposed to push the 2-D and 3-D features describing the same spatial patterns closer to each other while pushing the 2-D and 3-D features describing different spatial patterns farther away from each other. More specifically, $\mathcal{L}_{\text{ECL}}$ aligns the behavior-aware visual features e^v with the 3-D geometric-aware features e^g by constructing a common 2-D–3-D space for the visual encoder F_θ and the PCL encoder F_φ, as shown in Fig. 3. To this end, the multimodal contrastive learning loss is as follows:

$$\mathcal{L}_{\text{ECL}} = \mathcal{L}_{\text{cross}}\left(e^v, e^g\right) + \mathcal{L}_{\text{cross}}\left(e^g, e^v\right) \tag{4}$$

$$\mathcal{L}_{\text{cross}}\left(e^v, e^g\right) = \frac{1}{2N}\sum_{i=1}^{N} -\log \frac{\exp\left(\text{sim}\left(e_i^v, e_i^g\right)/\tau\right)}{\sum_{e^- \in e^g \setminus \{e_i^g\}} \exp\left(\text{sim}\left(e_i^v, e^-\right)/\tau\right)}. \tag{5}$$

$\mathcal{L}_{\text{ECL}}$ can not only transfer geometric information from 3-D to behavior-aware 2-D features but also transfer semantic details from 2-D to 3-D features. Overall, the visual encoder F_θ and the PCL encoder F_φ are trained by using the loss $\mathcal{L}_V = \mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{ECL}}$ and the loss $\mathcal{L}_{\text{ECL}}$, respectively. β is a weight used to balance the two losses. Moreover, F_θ is also trained by utilizing PPO as described in Section III-B. The visual encoder F_θ and PCL encoder F_φ pretrained by ECL will be integrated into the object discriminator to localize targets in Section III-E.

D. Dual-Thinking Scene Explorer

Following existing work [22], [23], [26], the explorer internally maintains a local 2-D semantic map $\mathcal{M}_{2D}^t$ for exploration decision-making. $\mathcal{M}_{2D}^t$ stores historical semantic priors and scene layouts, thereby enhancing navigation by mitigating the challenges of partial observability. $\mathcal{M}_{2D}^t$ is a $K \times X \times X$ matrix and $X \times X$ denotes the map size. Each element in this spatial map corresponds to a cell of size 25 cm^2 (5×5 cm) in the physical world. Each pixel on the egocentric top-down map is labeled with the corresponding semantic category, represented with a one-hot vector with $K = \mathcal{C} + 2$ channels, where $\mathcal{C}$ is the number of object categories. The extra channel indicates the obstacles and explored areas (see Appendix B for more details on 2-D semantic map construction).

During ObjectNav, the explorer should selectively access the frontiers in the local region where the target object is potentially present according to the scene priors in $\mathcal{M}_{2D}^t$. The frontier $f \in F$ refers to the critical region between explored and unexplored space, as shown in Fig. 2(b). Before the frontier selection, a utility function $\text{Score}(f) = U(f) - \alpha d_g(p_t, f)$ is employed to score all frontiers and assign them different access priorities. $U(f)$ denotes the information gain of the frontier f, which measures the free space left to explore beyond f, i.e., navigable cells unexplored in the partial semantic map [22], [30]. $d_g(p_t, f)$ denotes the geodesic distance between the agent's current location and the frontier f. The constant α adjusts the relative importance of both factors. By balancing information gain against distance cost, all frontiers in the map are sorted in descending order of score to assign different exploration priorities.

At the start of each ObjectNav episode, the semantic map is initialized with all zeros, $\mathcal{M}_0 = [0]^{K\times M\times M}$, which implies that there are no frontiers as explicit exploration goals. In this case, the explorer needs global guidance (Guid-G) that functions like a compass. As shown in Fig. 2(b), N_E global directional guides $\mathcal{A}_E \leftarrow [\mathcal{G}_E^i]_{i=0}^{N_E-1}$ are set up for the explorer π_φ ($N_E = 4$ in the illustration), where $\mathcal{A}_E$ denotes explorer's action space. As the task proceeds, once frontiers appear in the map and satisfy condition $d_g^{\text{thre}} < d_g(p_t, f) < d_g^{\max}/2$, the first N_E frontiers with the highest scores are taken as local exploration goals. $d_g^{\max}$ denotes the geodesic distance between the agent's current location and the farthest frontier. Empirically, our explorer considers only the closest half of all frontiers for each decision-making. Such a design prompts the explorer to thoughtfully focus on localized areas. d_g^{thre} denotes a lower bound threshold that prevents the agent from always greedily accessing the nearest frontier. Notably, the number of both global and local exploration goals is kept consistent, which facilitates the explorer to adaptively switch between fine-grained local exploration and holistic global exploration by switching the action space.

However, treasure hunting often does not go so well, just as the number of frontiers that satisfy the above condition will most likely be less than N_E. The variable number of local exploration actions poses a challenge to the dual-thinking explorer's training by using PPO. To tackle this challenge, we consider the existing $Q(0 < Q < N_E)$ frontiers as valid actions and utilize invalid action masking (IAM) [67] to mask out the logits corresponding to the $N_E - Q$ invalid actions during training. This is usually accomplished by replacing the logits of the actions to be masked in the PPO's policy gradient by a large negative number V (e.g., $V = -1 \times 10^8$) [67], [68]. Given state s_t, the process of IAM is considered as a differentiable function inv_s to be applied to the logits $l(s_t)$ outputted by exploration policy π_ϑ. The IAM process is defined as: $\pi_\vartheta(\cdot|s_t) = \text{softmax}(\text{inv}_s(l(s_t)))$, where

$$\text{inv}_s\left(l\left(s_t\right)\right)_i = \begin{cases} l_i, & \text{if } a_i \text{ is valid for } s_t \\ V, & \text{otherwise} \end{cases} \tag{6}$$

where l_i denotes the logit corresponding to the valid action a_i. The IAM process appears to make the gradient corresponding to the logits of the invalid action zero. In other words, taking invalid actions will not contribute to maximizing cumulative rewards; therefore, the explorer will tend not to take them. Once $Q = 0$, the explorer switches to global exploration mode to dispatch the agent to a new promising subregion. Otherwise, the explorer proceeds to visit the local frontiers.

As shown in Fig. 4, the dual-thinking explorer uses a map encoder F_ϑ and a PCL encoder F_φ pretrained by ECL to model the cumulative 2-D scene priors in $\mathcal{M}_{2D}^t$ and 3-D semantic

relations in $\mathcal{M}^t_{3D}$, respectively. In addition, a gated recurrent unit (GRU) [69] is used to summarize historical features, followed by an MLP to predict adaptive exploratory actions. Unlike the treasure-hunting task supported by a complete treasure map, the explorer does not know the semantic and layout clues of the scene in the initial phase of the ObjectNav episode. Therefore, a map encoder $F_\varphi = \text{Encoder}(\mathcal{M}^t_{2D})$ is employed to mine scene priors in the incrementally updated $\mathcal{M}^t_{2D}$ to model the mapping from the observation space to the action spaces. The map encoder is implemented as a standard UNet [70] encoder with four downsampling convolutional blocks.

E. Coarse-to-Fine Object Discriminator

The most intuitive object localization methods [20], [22], [23], [29] drive the agent to the corresponding locations as soon as the target semantic appears in the 2-D semantic map $\mathcal{M}^t_{2D}$. However, ambiguous semantics in $\mathcal{M}^t_{2D}$ are likely to mislead the agent to make wrong judgments, find the incorrect object, and finally terminate the task. Undeniably, although intuitive object localization is blind, it can indeed provide semantic inspiration for the agent to approach and confirm distant objects. Therefore, our insight is to consider this semantic inspiration as a precursor task to provide the agent with a proposal O^* of the object target. Subsequently, fine-grained 3-D object features are used to discriminate the proposal's authenticity. If the discrimination fails, the coarse-to-fine discriminator rejects the proposal and switches the agent to exploration mode until a new proposal is proposed.

Similar to the GC modeling, we also employ a 3-D semantic map $\mathcal{M}^t_{3D}$ to learn the object discriminator. Furthermore, the octree data structure in [71] is adopted to maintain the point cloud data, which helps to efficiently query the locations and semantics of the neighbors of any given point. Since we focus only on object identification, points with floor and wall semantics are filtered out, which significantly reduces storage and computation overhead, as shown in Fig. 4. In addition, a sliding window is utilized to capture the latest h-frame first-person view point clouds for constructing a local object map $\mathcal{M}^t_{3D}$. Our main point is that object recognition typically occurs only in the agent's current horizon or in a nearby local region. Given that there are Z objects in $\mathcal{M}^t_{3D}$, the object set is represented as $\mathbf{O} = \{O_m\}^Z_{m=1}$, in which O_m is depicted as the mth object. In each iteration of discrimination, $Z-1$ objects other than the proposal object O^*, i.e., $\{O_m \in \mathbf{O}\} \cap \{O_m \neq O^*\}$, are treated as the reference objects, which only provide the cues of locations or relations to the target object.

A recent study [72] suggests that embodied agents benefit from high-level 3-D scene representations to mitigate partial observability from limited FoV sensors. Therefore, the discriminator uses the visual encoder F_θ and PCL encoder F_φ pretrained in ECL as backbone networks to extract behaviorally aware visual features and 3-D geometric cues for discriminating the target proposal's authenticity, as shown in Fig. 4. In practice, F_θ and F_φ are implemented as the Mask-RCNN [64] with ResNet-50 [73] backbone and the PointNet [74], respectively. We assume that the criterion used by the discriminator to evaluate the proposal's authenticity is categorized into N_D levels: $\eta = \eta_{\text{low}} + \beta(1-\eta_{\text{low}})/N_D$, where $\eta_{\text{low}} = 0.5$ and discrimination action $\beta \in \mathcal{A}_D = \{0, 1, 2, \ldots, N_D - 1\}$. That is, the size of the discriminator's action space $\mathcal{A}_D$ is N_D. During the training of the discriminator, it learns the mapping from $\mathcal{M}^t_{3D}$ to $\mathcal{A}_D$ to set a threshold η for judging whether the proposal is true or not. The octree-based point cloud organization facilitates us to query the semantics $\hat{P}^t_{s,i}$ of the eight neighbors $\{\hat{P}^t_i\}_{0\le i\le 7}$ of each point in the proposal O^*. Once the probabilities of a point and more than half of its neighbors belonging to the target semantic category are all greater than η, i.e., $P^t_s[n_\mathcal{C}] > \eta$ and $\sum_{0\le i\le 7} \mathbb{I}\{\hat{P}^t_{s,i}[n_\mathcal{C}] > \eta\} \ge 4$, the proposal is recognized as the object target. Here, $n_\mathcal{C}$ denotes the dimension index corresponding to the target semantic $\mathcal{C}$. Accordingly, the nearest one to the agent among all such points in the proposal O^* is used as the agent's current navigation goal.

Notably, the discrimination is performed in real time during the agent's travel toward the proposal, rather than on a one-time basis. The continuous discrimination ensures the robustness of object localization. If the distance between the agent and the proposal is less than 1 m and the discrimination is true, the ObjectNav succeeds. Otherwise, the semantic of the region where the proposal is located is eliminated to prevent it from misleading the agent to find the wrong object instance. Semantic elimination means setting the corresponding locations in the 2-D semantic map to zero, i.e., the object proposal will be neglected by the agent.

F. Reward Shapping and Implementation Details of the ObjectNav Policy

1) Rewards: Since the explorer and the discriminator have their own responsibilities, we design different reward functions for them. For the discriminator, when it eliminates the semantic of a proposal, the reward is $R_D = -10$ if the geodesic distance between the agent and the object target is less than 1 m (i.e., $d_g(p_t, \text{target}) < d_s$), and $R_D = 5$ if $d_g(p_t, \text{target}) > d_s$. $d_g(\cdot,\cdot)$ can be queried from the habitat simulator during training. If the agent successfully recognizes and navigates to the object target, then $R_D = 15$. Both the explorer and the discriminator are trained using the PPO algorithm.

The explorer's reward function consists of three parts.

1) A time penalty $\lambda = -0.001$ at each timestep t, which is used to encourage faster ObjectNav.
2) An immediate distance reward $R^d_E \propto (d_g(p_t, \text{target}) - d_g(p_{t-1}, \text{target}))$, which is used to motivate the agent to approach the object target.
3) A novelty reward $R^n_E \propto (r^n_t - r^n_{t-1})$ [27], which is used to motivate the explorer to explore previously unvisited areas, where $r^n_t = \sum_{\mathbf{A}_i \in \mathbf{S}_t} \mathbf{A}_i / \sqrt{n_i}$. $\mathbf{S}_t$ represents all the cells in the 2-D semantic map at time t, and $\mathbf{A}_i$ is the area explored by the agent in the grid cell i. During exploration, each cell is assigned an access count n_i, which records the number of times the grid i has been observed. Therefore, the explorer's reward function can be denoted as $R_E = \lambda + R^d_E + R^n_E$.

2) Local Deterministic Strategy: The initial phase of ObjectNav is dominated by the explorer. Until the object

target's semantics emerge in the semantic map, the discriminator takes over the searching procedure and travels to the suspected target for discrimination. During each episode, exploration and discrimination run alternately. The explorer provides the agent with local frontier goals or globally steered goals, and the discriminator offers proposal goals to the agent. The fast marching method [75] is used to plan the shortest path from the agent's current location to these goals, which the agent follows by taking deterministic actions. Empirically, the explorer samples a new goal every $u = 20$ timesteps, which prevents the explorer from deadlocking caused by frequent decision-making.

IV. Experiments

A. Experimental Setups

1) ObjDet and InstSeg: We perform experiments on the Habitat simulator [76] with the Gibson [32] dataset that contains photorealistic 3-D reconstructions of real-world environments. Following previous works [20], [51], we use 25 train/5 val scenes from the Gibson tiny split for our experiments, where semantic annotations are available. The agent's observation space consists of 640 × 480 RGB-D images. The agent's discrete action space consists of Move_Forward (0.25m), Turn_Left (30°), and Turn_Right (30°). Our agent performs about 819k frames of interactive learning in diverse scenarios to optimize the exploration policy E^2-CL, the visual encoder f_θ, and the PCL encoder f_φ. Our ECL is carried out on four NVIDIA 3090 GPUs and takes about 72 h. The pseudocode, hyperparameters, and training curves of ECL are listed in Appendix C.

For downstream task retraining, our agent actively acts to collect 10k visual frames in each training scene using E^2-CL. That is, a total of 250k visual frames are collected. Meanwhile, the corresponding semantic labels are extracted from the Habitat simulator [76] as supervision signals for retraining ObjDet and InstSeg models. Following evaluation setups in previous works [20], [51], we use six common indoor object categories for our experiments: chair, couch, bed, toilet, TV, and potted plant. Thus, 250k frames of images contain a total of about 316k objects. We consider two evaluation settings.

1) *Train Split:* The 5k images containing 6393 objects are randomly sampled from 25 in-distribution training scenes (200 images per scene).
2) *Test Split:* The 5k images containing 5025 objects are randomly sampled from 5 out-of-distribution test scenes (1000 images per scene).

Our method is implemented by using the collected 250k frames to retrain the ResNet50 pretrained by ECL. To highlight the contributions of our ECL, the following baselines are selected for comparative study: SimCLR [77], CRL [43], OVRL [54], Ego^2-MAP† [57], Pri3D [60], and MIT [61]. Among them, CRL [43] employs an adversarial contrastive loss for embodied visual representation learning based on only RGB images. Ego^2-MAP [57] proposes a multimodal contrastive representation learning method based on visual features and 2-D semantic maps. Since we do not have access to the source code of Ego^2-MAP, the RGB images and the 2-D semantic maps in our method are utilized to replicate the approach as much as possible. The replicated approach is named Ego^2-MAP†. To ensure fair comparisons, the RGB images used to pretrain our method are saved for the pretraining of SimCLR [77] and OVRL [54]. Besides the RGB images, additional 2-D semantic maps are collected for the pretraining of Ego^2-MAP†. Moreover, another two fundamental baselines are set: 1) a raw ResNet50 without pretraining is trained in a supervised manner using the collected 250k frames, which is named From Scratch; and 2) the collected 250k frames are used to retrain the ImageNet-supervised pretrained weights, which is named ImageNet Supervised.

We report the bounding box and the mask AP scores for ObjDet and InstSeg tasks, respectively. AP is the average precision averaged over multiple intersection over union (IoU) (10 IoU thresholds of 0.50:0.05:0.95) and is the primary challenge metric in the COCO dataset [78]. IoU is defined to be the IoU of the predicted and ground-truth bounding box or the segmentation mask.

2) Object Navigation: We perform experiments on the MP3D [31], Gibson [32], and HM3D [33] datasets with the Habitat simulator [76]. For Gibson, we use 25 train/5 val scenes from the Gibson tiny split. Following existing works [26], we consider six goal categories, including chair, couch, potted plant, bed, toilet, and TV. For HM3D, we use the standard splits of 75 train/20 val scenes with habitat format and consider the same goal categories as Gibson. For MP3D, we use the standard split of 61 train/11 val scenes with the Habitat ObjectNav dataset [76], which consists of 21 goal categories.

In the pretraining phase, the visual encoder F_θ and the PCL encoder F_φ are optimized using our ECL in diverse Gibson, MP3D, and HM3D scenes. Subsequently, F_θ and F_φ are integrated into the framework shown in Fig. 4 for ObjectNav. Both our ECL and ObjectNav agents have 640 × 480 RGB-D observation spaces for constructing semantic maps. The discrete action space of our ObjectNav agent consists of Move_Forward (0.25 m), Turn_Left (30°), Turn_Right (30°), and Stop. Note that the RGB-D $\{o_t, d_t\}$ and pose s_t readings are noise-free from simulation. The pretrained 2-D semantic model RedNet [66] and the Mask-RCNN [64] trained with the COCO dataset [78] are employed for 2-D and 3-D semantic mapping on all three datasets. For each frame, we randomly sample 512 points for PCL-based 3-D semantic construction. That is, the number of points in the 3-D local semantic map is $Q^t = 512 \times (l + 1)$. During training, we sample actions every 25 steps and use the PPO [65] for both object localization (discriminator) and scene exploration (explorer) subpolicies. We adopt the Adam optimizer with a learning rate of 0.000025 and a discount factor of 0.99. The size of the semantic metric map $\mathcal{M}^t_{2D}$ is 48 × 48 m and the map resolution is 5 cm. We train the model for over 2 million frames on four NVIDIA 3090 GPUs. If not specified, the following parameter settings are used: $\alpha = 0.2$, $h = 16$, $N_E = 4$, $d_g^{\text{thre}} = 2$m, and $N_D = 14$.

The following three metrics are reported to evaluate all methods quantitatively.

1) *Success Rate (SR):* Percentage of successful episodes.

TABLE I

PERFORMANCE OF DIFFERENT METHODS ON OBJDET AND INSTSEG TASKS

Method (Venue)	Train Split		Test Split	
	ObjDet	InstSeg	ObjDet	InstSeg
From Scratch	62.93	52.94	15.01	13.20
ImageNet Supervised	70.03	60.47	22.20	19.43
SimCLR [77] (ICML 2020)	66.37	54.82	20.17	18.40
CRL [43] (ICCV 2021)	68.21	56.54	22.45	19.61
Pri3D [60] (ICCV 2021)	70.28	59.86	25.34	23.04
OVRL [54], [55] (ICLR 2023)	67.56	54.82	22.23	20.18
Ego^2-MAP† [57] (ICCV 2023)	67.29	54.97	20.72	19.89
MIT [61] (ICCV 2023)	68.30	56.88	24.19	22.61
From ECL (Ours)	**72.32**	**62.11**	**25.89**	**23.81**

TABLE II

OBJECTNAV RESULTS ON MP3D (VAL). † DENOTES THE RESULTS WE OBTAINED USING THE OFFICIAL OPEN-SOURCE CODE

Method (Venue)	MP3D (val)			
	SR (%) ↑	SPL (%) ↑	DTS (m) ↓	Ext. Data/Model
DD-PPO [79] (ICLR 2019)	8.0	1.8	6.90	no
FBE [81] (First proposed in 1997)	22.7	7.2	6.70	no
ANS [85] (CVPR 2020)	21.2	9.4	6.30	no
SemExp [20] (NeurIPS 2020)	28.3	10.9	6.06	no
Red-Rabbit [9] (ICCV 2021)	34.6	7.9	-	no
TreasureHunt [7] (ICCV 2021)	28.4	11.0	5.58	yes
Habitat-Web [8] (CVPR 2022)	35.4	10.2	-	yes
L2M [21] (ICLR 2022)	32.1	11.0	5.12	no
PONI [22] (CVPR 2022)	27.8	12.0	5.60	no
OVRL [54] (ICLR 2023)	28.6	7.4	-	no
Ego^2-MAP [57] (ICCV 2023)	29.0	10.6	5.17	yes
3D-Aware† [26] (CVPR 2023)	33.4	13.6	5.03	no
SGM [86] (CVPR 2024)	37.7	14.7	4.93	GPT-4 [92]
CER [24] (TCSVT 2024)	37.2	14.9	5.01	no
HOP_i++ [80] (TPAMI 2025)	32.1	11.4	5.32	no
HOP_e++ [80] (TPAMI 2025)	37.0	15.2	**4.11**	no
ECL-ObjectNav (Ours)	**37.8**	**15.5**	4.95	no

2) *SPL:* Success weighted by path length, which measures the efficiency of the agent over oracle path length.
3) *DTS:* Geodesic distance of agent to the object goal at the end of the episode.

In addition, we report which methods use external data or model (Ext. Data/Model) to enhance ObjectNav. Besides using the Random Sample and classical FBE as nonlearning baselines, we consider the following mainstream baselines in the ObjectNav task.

1) *End-to-End Strategies:* DD-PPO [79], Habitat-Web [8], Red-Rabbit [9], TreasureHunt [7], OVRL [54], Ego^2-MAP [57], and HOP_i++ [80].
2) *Modular Strategies:* FBE [81], ANS [82], SemExp [20], FSE [30], L2M [21], PONI [22], 3-D-Aware [26], L3MVM [83], ESC [84], Pixel-Nav [85], HOP_i++ [80], SGM [86], VoroNav [87], and ImagineNav [88].

B. Comparative Studies for ObjDet and InstSeg

The quantitative comparative results between our method and several baselines on the object recognition tasks are shown in Table I. First, as expected, both self-supervised and Imagenet Supervised baselines have significant performance gains relative to From Scratch. This phenomenon reflects that different types of pretraining can improve the models' performance on both tasks. Second, our method and Pri3D outperform Imagenet Supervised pretraining models on both tasks. These results reflect the potential and superiority of 2-D–3-D self-supervised contrastive learning, i.e., the 3-D scene priors can enrich the 2-D visual features. Third, the advantages of ECL over offline self-supervised learning are revealed by comparing CRL with SimCLR (or OVRL). Notably, CRL, SimCLR, and OVRL are retrained using the same 250k visual frames. The difference is that CRL uses an adversarial contrastive loss for interactive visual pretraining, while SimCLR and OVRL are pretrained offline using saved data.

Fourth, the experimental results of Ego^2-Map† indicate that contrastive learning between 2-D visual features and 2-D semantic maps yields no significant performance improvement. In contrast, the exchange of information between 2-D visual features and 3-D geometric cues facilitates the accuracy of object recognition tasks. In particular, Pri3D improves the AP metrics by 3.91 (Train Split) and 5.17 (Test Split) absolutely on the ObjDet task relative to the self-supervised learning baseline (SimCLR). Similarly, Pri3D improves the AP metrics by 5.04 (Train Split) and 4.64 (Test Split) absolutely on the InstSeg task relative to the self-supervised learning baseline (SimCLR). Finally, our method achieves the best AP metrics on two splits of both tasks. On the one hand, by comparing our method with offline self-supervised learning baselines (methods other than CRL), the results reveal the necessity of ECL using 2-D visual features and 3-D geometric structures. On the other hand, by comparing it with existing offline 2-D–3-D contrastive representation learning baselines (Pri3D and MIT), our results demonstrate the superiority of our BA and GC modeling.

C. Comparative Studies for ObjcetNav

Our ECL-enhanced ObjectNav policy is evaluated on the MP3D (val), Gibson (val), and HM3D (val) datasets. As shown in Table II, our method improves the SR and SPL metrics by 15.1%~29.8% and 8.3%~13.7% compared to the learning-based (DD-PPO) and classical frontier-based (FBE) baselines, respectively. Our policy outperforms existing SOTA end-to-end (OVRL and HOP_i++) and modular (3-D-Aware†, SGM, and HOP_e++) approaches in terms of SR and SPL metrics. In addition, our scheme achieves a competitive DTS metric on MP3D (val). Notably, Habitat-Web and SGM achieved competitive and suboptimal SR metrics on MP3D (Val) thanks to the use of a large amount of external data and an external large language model, respectively. Nevertheless, our scheme outperforms the external data/model enhanced methods (TreasureHunt, Habitat-Web, Ego^2-MAP, and SGM).

As an end-to-end approach, OVRL uses a knowledge distillation-like self-supervised learning method to pretrain the visual encoder. In contrast, Ego^2-MAP further introduces additional contrastive pretraining based on 2-D visual features and 2-D scene priors. However, OVRL and Ego^2-MAP perform less well than our method. The superiority of our approach is partly attributed to modeling BA and GC in an embodied manner and partly to a modular strategy design based on semantic maps. More significantly, our approach likewise outperforms the modular and semantic map-based ObjectNav schemes (PONI, 3-D-Aware†, and HOP_e++). In particular, our approach outperforms 3-D-Aware† and HOP_e++, which attempts to exploit 3-D scene priors and hierarchical scene priors, respectively. Such experimental results reflect the contributions of our ECL-based pretrained point cloud encoder

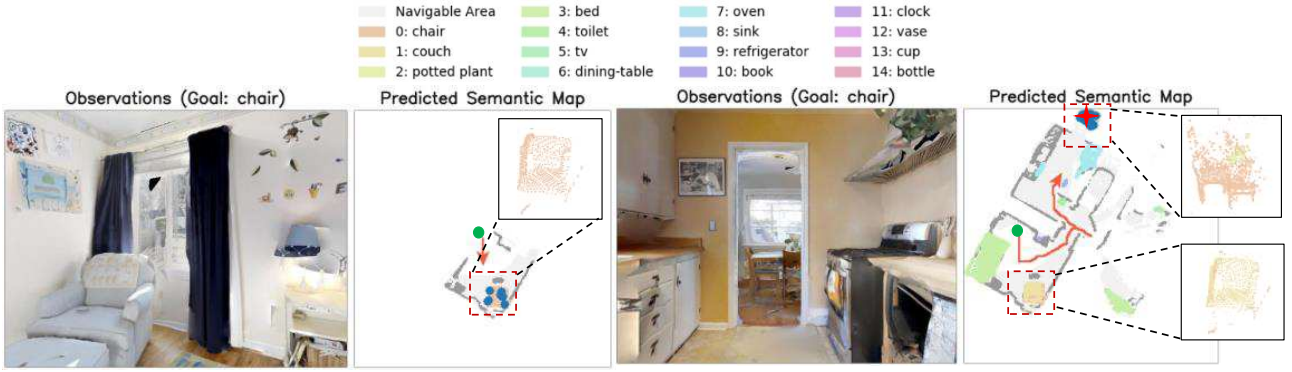

Fig. 5. ObjectNav demo on the Gibson dataset.

TABLE III

OBJECTNAV RESULTS ON GIBSON (VAL). † DENOTES THE RESULTS WE OBTAINED USING THE OFFICIAL OPEN-SOURCE CODE

Method (Venue)	Gibson (val)			
	SR (%) ↑	SPL (%) ↑	DTS (m) ↓	Ext. Data/Model
DD-PPO [79] (ICLR 2019)	15.0	10.7	3.240	no
FBE [81] (First proposed in 1997)	41.7	21.4	2.63	no
Random Sample	54.4	28.8	1.92	no
ANS [82] (CVPR 2020)	67.1	34.9	1.66	no
SemExp [20] (NeurIPS 2020)	65.2	33.6	1.52	no
PONI [22] (CVPR 2022)	73.6	41.0	1.25	no
FSE [30] (ICRA 2023)	71.5	36.0	1.35	no
3D-Aware† [26] (CVPR 2023)	73.8	39.6	1.39	no
ECL-ObjectNav (Ours)	**74.6**	**41.7**	**1.27**	no

TABLE IV

OBJECTNAV RESULTS ON HM3D (VAL)

Method (Venue)	HM3D (val)		
	SR (%) ↑	SPL (%) ↑	Ext. Data/Model
DD-PPO [79] (ICLR 2019)	26.0	12.0	no
FBE [81] (First proposed in 1997)	23.7	12.3	no
SemExp [20] (NeurIPS 2020)	37.9	18.8	no
Habitat-Web [8] (CVPR 2022)	55.0	22.0	yes
FSE [30] (ICRA 2023)	53.8	24.6	no
L3MVM [83] (IROS 2023)	54.2	25.5	GPT-2-Large [93]
ESC [84] (ICML 2023)	39.2	23.3	GPT-3.5 [94]
Pixel-Nav [85] (ICRA 2024)	37.9	20.0	GPT-4 [92]
VoroNav [87] (ICML 2024)	46.5	25.7	GPT-4 [92]
ImagineNav [88] (ICLR 2025)	53.0	23.8	GPT-4o-mini [92]
ECL-ObjectNav (Ours)	**55.8**	**27.9**	no

and visual encoder for scenario exploration and object recognition tasks. Moreover, our BA and GC-based embodied pretraining provides new ways for active scene perception and object recognition.

As shown in Table III, our method's performance on Gibson (val) is similar to that on MP3D (val). These experimental results reflect that our scheme works universally on different datasets. Quantitatively, our method improves the SR and SPL metrics by 0.8%~3.1% and 0.7%~5.7% compared to the SOTA methods (FSE, PONI, and 3-D-Aware†), respectively. Our coarse-to-fine ObjectNav strategy follows the exploration–discrimination–exploitation framework, which is distinct from that of SemExp, FSE, and 3-D-Aware†. This is one of the reasons why we achieve good ObjectNav performance. Qualitatively, a specific ObjectNav example on the Gibson dataset is shown in Fig. 5. The agent initially recognizes the couch as a chair due to a semantic segmentation error. Although the segmentation error has been corrected to some extent, point clouds with the couch semantic in the map

TABLE V

ABLATION STUDIES OF SPECIFIC COMPONENTS IN ECL

Ablations				Train Split		Test Split	
$\mathcal{R}_V$	$\mathcal{R}_A$	$\mathcal{L}_{CE}$	$\mathcal{L}_{ECL}$	ObjDet	InstSeg	ObjDet	InstSeg
✓	✓			67.59	55.27	21.62	20.08
✓	✓		✓	70.10	59.88	23.65	22.92
✓	✓	✓		70.62	60.19	23.43	22.38
✓	✓	✓	✓	**72.32**	**62.11**	**25.89**	**23.81**
		✓	✓	67.89	56.03	21.74	20.81
✓		✓	✓	69.33	58.94	24.10	21.25
	✓	✓	✓	70.51	59.26	25.22	23.47

TABLE VI

ABLATION STUDIES OF SPECIFIC FEATURES IN OBJECTNAV

Ablations		Gibson (val)		
F_θ	F_φ	SR (%) ↑	SPL (%) ↑	DTS (m) ↓
✓		74.4	41.5	1.28
	✓	74.0	41.2	1.30
✓	✓	**74.6**	**41.7**	**1.27**

are still intermingled with point clouds with chair semantics. Luckily, the features provided by our pretrained visual encoder motivate the agent to make correct recognition and navigate to a chair in another room.

As shown in Table IV, benefiting from the coarse-to-fine treasure-hunting mindset and the cooperation of explorer and discriminator, our method, respectively, improves the SR and SPL metrics by 0.8%~17.9% and 2.2%~9.1% on HM3D (Val) relative to the SOTA methods. Notably, L3MVM, ESC, Pixel-Nav, and VoroNav fully exploit the domestic scene layouts and semantic priors internalized in the large language model to solve the ObjectNav problem on the val split without learning from the training split. Despite the attempts of different knowledge extraction techniques, our method outperforms them in terms of SR and SPL metrics. In addition, our method outperforms the latest ObjectNav solution (ImagineNav) based on a multimodal large model. An example of our ObjectNav policy is shown in Fig. 6, which adequately reflects the cooperation between the explorer and discriminator. More ObjectNav demos can be found in Appendix E.

D. Ablation Studies

1) Do the Individual Modules in the ECL Work?: We conduct ablation studies on specific components in the ECL, and the results are shown in Table V. Specifically, we employ combinations of different components to implement ECL and evaluate the pretrained models on two object recognition tasks.

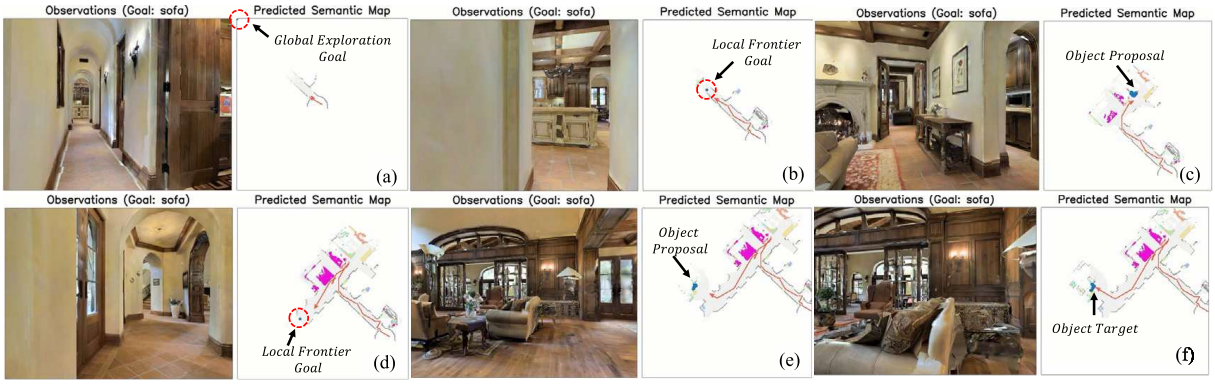

Fig. 6. (a) Task begins, the agent is dominated by a global exploration goal to fully explore the unknown scene. (b) Explorer selects a frontier where the object target is most likely to be found based on the semantic map. (c) Due to a semantic segmentation error, the agent initiates a proposal by mistaking a chair for a sofa. (d) Discriminator eliminates the semantic ambiguity caused by the chair, while the explorer picks another optimal frontier. (e) Sofa's semantic reappears in the 2-D semantic map. The agent is guided by the 2-D semantic inspiration to approach the proposal and identify its authenticity in real time. (f) Discrimination succeeds and the agent is close enough to the proposal; thus, the ObjectNav task succeeds.

TABLE VII

ABLATION STUDIES OF DIFFERENT COMPONENTS OF THE COARSE-TO-FINE OBJECTNAV POLICY

Ablations						MP3D (Val)	
R_E^d	R_E^n	Guid-F	Guid-G	Proposal-2D	Recog-3D	SR (%) ↑	SPL (%) ↑
✓	✓	✓	✗	✓	✓	36.3	14.5
✓	✓	✗	✓	✓	✓	36.0	14.1
✓	✓	✓	✓	✓	✗	17.2	8.0
✓	✓	✓	✓	✗	✓	35.7	13.6
✗	✓	✓	✓	✓	✓	34.8	13.5
✓	✗	✓	✓	✓	✓	36.1	14.0
✓	✓	✓	✓	✓	✓	**37.8**	**15.5**

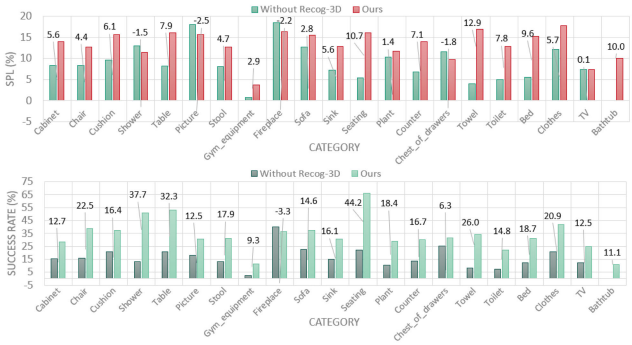

Fig. 7. SR and SPL metrics of each category when evaluated on the MP3D dataset. Recog-3-D represents the discriminator's 3-D object map-based object recognition. The black numbers indicate relative performance gains.

We first ablate BA $\mathcal{L}_{\text{CE}}$ and GC $\mathcal{L}_{\text{ECL}}$ while retaining the exploration policy E^2-CL ($\mathcal{R}_V$ and $\mathcal{R}_A$) (Lines 1–3). The results reflect that both BA and GC enhance our method's performance. The best results are achieved when both are used at the same time. In addition, we retain BA and GC and ablate the different components of E^2-CL. Line 5 of Table V indicates that a randomized wandering exploration method is used. The results show that both $\mathcal{R}_V$ and $\mathcal{R}_A$ enhance our method. Notably, Line 3 of Table V shows that better AP metrics can also be achieved by using only $\mathcal{L}_{\text{CE}}$ (without $\mathcal{L}_{\text{ECL}}$), which suggests that action awareness facilitates the agent's active movement and perception in the scenarios. The comparison with the randomized wandering exploration also demonstrates the superiority of our E^2-CL. As shown in Table VI, the contributions of F_θ and F_φ are ablated by adopting or not adopting the pretrained models to initialize the ObjectNav policy. The results show the enhanced ObjectNav performance by using our BA and GC-based 2-D–3-D multimodal ECL. Both pretrained F_θ and F_φ can boost the SR and SPL metrics.

2) Do the Different Stages of Explorer and Discriminator Work?: Lines 1 and 2 of Table VII ablate the explorer's Guid-F and Guid-G, respectively. Quantitatively, the absence of Guid-F and Guid-G decreases the SR metric by 1.5% and 1.8% relative to the full method (Line 7 of Table VII), respectively. Correspondingly, the SPL metrics also decreased. Experimental results show that both Guid-F and Guid-G enhance the ObjectNav performance, but Guid-F contributes more.

Lines 3 and 4 of Table VII ablate the discriminator's 2-D object proposal (Proposal-2D) and 3-D object recognition (Recog-3D), respectively. Surprisingly, the absence of Recog-3D means the removal of the discriminative power, leading to a substantial drop in both SR and SPL metrics. The ablation of Recog-3D reflects the misleading effect of ambiguous semantic segmentation on agents. Fig. 7 shows the SR and SPL metrics for each category with and without 3-D object recognition when evaluated on the MP3D dataset. Significantly, the discrimination–exploitation-based object localization improves the SR and SPL metrics for almost all categories, which reflects our discriminator's strength. The absence of Recog-2D means the lack of coarse-grained 2-D semantic inspiration. Nonetheless, with the explorer leading the way, the agent can still localize object targets well.

3) Does the Reward Shaping Work?: Lines 5 and 6 of Table VII ablate the explorer's immediate distance reward R_E^d and novelty reward R_E^n, respectively. The experimental results reflect the importance of R_E^d for reward shaping, i.e., the absence of R_E^d results in 3.0% and 2.0% decreases in SR and SPL metrics, respectively. R_E^d alleviates the reward sparsity problem during strategy learning by motivating the agent to approach the object target through an immediate bonus. R_E^n has a relatively smaller impact on the agent but also contributes significantly to the ObjectNav performance.

See Appendix D for additional diagnostic studies and discussion. More discussions on technical details can be found in Appendix F.

V. Conclusion and Future Work

This article proposes an ECL-enhanced coarse-to-fine ObjectNav policy with explorer and discriminator cooperation to tackle the ObjectNav problem in previously unknown indoor settings. Specifically, we first propose a novel ECL method based on BA and GC, which is equipped with an exploration strategy E^2-CL based on a curiosity and action awareness-driven contrastive reward. Unlike previous self-supervised learning methods based on RGB images, our approach performs an organic cross-modal fusion of semantically rich 2-D visual patterns and 3-D geometric structural features. Our embodied BA and GC modeling outperforms existing purely vision-based self-supervised learning approaches and offline 2-D–3-D contrastive representation learning techniques on object recognition tasks.

Furthermore, a modular ObjectNav strategy with explorer and discriminator cooperation is proposed under the exploration–discrimination–exploitation framework. The dual-thinking explorer can instantly switch its exploratory mindset depending on current observations to prevent inefficient exploration, just as it can skillfully employ the local and Guid-G, respectively, provided by a treasure map and a compass during treasure hunting. The coarse-to-fine discriminator can identify object targets using 3-D geometric and semantic patterns to prevent blind object localization. By integrating the PCL encoder and visual encoder pretrained by ECL into the explorer and discriminator, our policy achieves the best ObjectNav performance on the MP3D, Gibson, and HM3D datasets. These improvements are attributed to the adequate scene representation capability and excellent object recognition potential of our pretrained models. In particular, the ablation studies also indicate the effective contributions of the individual components in our method.

In the future, more efficient embodied learning paradigms and 3-D feature extraction methods need to be further investigated and discussed. This work emphasizes the coarse-to-fine thinking of object seeking, which implies that the design and implementation of the explorer and discriminator can be further optimized. We place this work in the future, e.g., improving object recognition accuracy by mining co-occurrence relationships among 3-D objects.

Acknowledgment

This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

References

[1] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Sci. Robot.*, vol. 8, no. 79, p. 6991, Jun. 2023.

[2] D. Batra et al., "ObjectNav revisited: On evaluation of embodied agents navigating to objects," 2020, *arXiv:2006.13171*.

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

[4] H. Du, X. Yu, and L. Zheng, "Learning object relation graph and tentative policy for visual navigation," in *Eur. Conf. Comput. Vis.*, 2020, pp. 19–34.

[5] S. Zhang, X. Song, Y. Bai, W. Li, Y. Chu, and S. Jiang, "Hierarchical object-to-zone graph for object navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15130–15140.

[6] K. Zhou, C. Guo, W. Guo, and H. Zhang, "Learning heterogeneous relation graph and value regularization policy for visual navigation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16901–16915, Nov. 2024.

[7] O. Maksymets et al., "THDA: Treasure hunt data augmentation for semantic navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 15374–15383.

[8] R. Ramrakhya, E. Undersander, D. Batra, and A. Das, "Habitat-Web: Learning embodied object-search strategies from human demonstrations at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5173–5183.

[9] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary tasks and exploration enable objectgoal navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16117–16126.

[10] W. Li, R. Hong, J. Shen, L. Yuan, and Y. Lu, "Transformer memory for interactive visual navigation in cluttered environments," *IEEE Robot. Autom. Lett.*, vol. 8, no. 3, pp. 1731–1738, Mar. 2023.

[11] B. Mayo, T. Hazan, and A. Tal, "Visual navigation with spatial attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16898–16907.

[12] C. Zhao, Y. Qi, and Q. Wu, "Mind the gap: Improving success rate of vision-and-language navigation by revisiting oracle success routes," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 4349–4358.

[13] E. C. Tolman, "Cognitive maps in rats and men," *Psychol. Rev.*, vol. 55, no. 4, pp. 189–208, 1948.

[14] J. O'Keefe and N. Burgess, "Geometric determinants of the place fields of hippocampal neurons," *Nature*, vol. 381, no. 6581, pp. 425–428, May 1996.

[15] H. Zeng, X. Song, and S. Jiang, "Multi-object navigation using potential target position policy function," *IEEE Trans. Image Process.*, vol. 32, pp. 2608–2619, 2023.

[16] G. Kumar, N. S. Shankar, H. Didwania, R. D. Roychoudhury, B. Bhowmick, and K. M. Krishna, "GCExp: Goal-conditioned exploration for object goal navigation," in *Proc. 30th IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2021, pp. 123–130.

[17] R. Dang et al., "Search for or navigate to? Dual adaptive thinking for object navigation," in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, pp. 8250–8259, Oct. 2023.

[18] W. Li, X. Song, Y. Bai, S. Zhang, and S. Jiang, "ION: Instance-level object navigation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4343–4352.

[19] R. Dang, Z. Shi, L. Wang, Z. He, C. Liu, and Q. Chen, "Unbiased directed object attention graph for object navigation," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3617–3627.

[20] D. S. Chaplot, D. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 4247–4258.

[21] G. Georgakis, B. Bucher, K. Schmeckpeper, S. Singh, and K. Daniilidis, "Learning to map for active semantic goal navigation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=swrMQttr6wN

[22] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "PONI: Potential functions for ObjectGoal navigation with interaction-free learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18890–18900.

[23] A. J. Zhai and S. Wang, "PEANUT: Predicting and navigating to unseen targets," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10892–10901.

[24] B. Chen et al., "Think holistically, act down-to-Earth: A semantic navigation strategy with continuous environmental representation and multi-step forward planning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3860–3875, May 2024.

[25] H. Luo, A. Yue, Z.-W. Hong, and P. Agrawal, "Stubborn: A strong baseline for indoor object navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 3287–3293.

[26] J. Zhang et al., "3D-aware object goal navigation via simultaneous exploration and identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6672–6682.

[27] S. K. Ramakrishnan, D. Jayaraman, and K. Grauman, "An exploration of embodied visual exploration," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1616–1649, May 2021.

[28] H. Zhu et al., "EXCALIBUR: Encouraging and evaluating embodied exploration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14931–14942.

[29] Y. Liang, B. Chen, and S. Song, "SSCNav: Confidence-aware semantic scene completion for visual semantic navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13194–13200.

[30] B. Yu, H. Kasaei, and M. Cao, "Frontier semantic exploration for visual target navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4099–4105.

[31] A. Chang et al., "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 667–676.

[32] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson Env: Real-world perception for embodied agents," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9068–9079.

[33] S. K. Ramakrishnan et al., "Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI," 2021, *arXiv:2109.08238*.

[34] M. Z. Irshad, N. C. Mithun, Z. Seymour, H.-P. Chiu, S. Samarasekera, and R. Kumar, "Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 4065–4071.

[35] S. Chen, P.-L. Guhur, C. Schmid, and I. Laptev, "History aware multimodal transformer for vision-and-language navigation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 5834–5847.

[36] C. Lin, Y. Jiang, J. Cai, L. Qu, G. Haffari, and Z. Yuan, "Multimodal transformer with variable-length memory for vision-and-language navigation," in *Proc. Eur. Conf. Comput. Vis.*, 2021, pp. 380–397.

[37] Y. Zhao et al., "Target-driven structured transformer planner for vision-language navigation," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4194–4203.

[38] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Think global, act local: Dual-scale graph transformer for vision-and-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16516–16526.

[39] C. Gao et al., "Adaptive zone-aware hierarchical planner for vision-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14911–14920.

[40] X. Li, Z. Wang, J. Yang, Y. Wang, and S. Jiang, "KERM: Knowledge enhanced reasoning for vision-and-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2583–2592.

[41] H. Wang, W. Liang, L. Van Gool, and W. Wang, "Dreamwalker: Mental planning for continuous vision-language navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10873–10883.

[42] S. Zhang, X. Song, W. Li, Y. Bai, X. Yu, and S. Jiang, "Layout-based causal inference for object navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10792–10802.

[43] Y. Du, C. Gan, and P. Isola, "Curious representation learning for embodied intelligence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10388–10397.

[44] J. F. Henriques and A. Vedaldi, "MapNet: An allocentric spatial memory for mapping environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8476–8484.

[45] V. Cartillier, Z. Ren, N. Jain, S. Lee, I. Essa, and D. Batra, "Semantic MapNet: Building allocentric semantic maps and representations from egocentric views," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 964–972.

[46] S. K. Ramakrishnan, Z. Al-Halah, and K. Grauman, "Occupancy anticipation for efficient exploration and navigation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 400–418.

[47] P. Zhong, B. Chen, S. Lu, X. Meng, and Y. Liang, "Information-driven fast marching autonomous exploration with aerial robots," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 810–817, Apr. 2022.

[48] B. Chen, Y. Cui, P. Zhong, W. Yang, Y. Liang, and J. Wang, "STExplorer: A hierarchical autonomous exploration strategy with spatio-temporal awareness for aerial robots," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 6, pp. 1–24, Dec. 2023.

[49] B. Chen, S. Lu, P. Zhong, Y. Cui, Y. Liang, and J. Wang, "SemNav-HRO: A target-driven semantic navigation strategy with human–robot–object ternary fusion," *Eng. Appl. Artif. Intell.*, vol. 127, Jan. 2024, Art. no. 107370.

[50] B. Chen et al., "Socially aware object goal navigation with heterogeneous scene representation learning," *IEEE Robot. Autom. Lett.*, vol. 9, no. 8, pp. 6792–6799, Aug. 2024.

[51] D. S. Chaplot, M. Dalal, S. Gupta, J. Malik, and R. Salakhutdinov, "SEAL: Self-supervised embodied active learning using exploration and 3D consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, Mali, 2021, pp. 13086–13098.

[52] R. Liu, X. Wang, W. Wang, and Y. Yang, "Bird's-eye-view scene graph for vision-language navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 10968–10980.

[53] D. An et al., "BEVBert: Multimodal map pre-training for language-guided navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2022, pp. 2737–2748.

[54] K. Yadav et al., "Offline visual representation learning for embodied navigation," in *Proc. Workshop Reincarnating Reinforcement Learn. (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=GeJckQWZlMX

[55] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[56] K. P. Singh, J. Salvador, L. Weihs, and A. Kembhavi, "Scene graph contrastive learning for embodied navigation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10850–10860.

[57] Y. Hong et al., "Learning navigational visual representations with semantic map supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3032–3044.

[58] P. Arsomngern, S. Nutanong, and S. Suwajanakorn, "Learning geometric-aware properties in 2D representation using lightweight CAD models, or zero real 3D pairs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21371–21381.

[59] N. Chen, L. Chu, H. Pan, Y. Lu, and W. Wang, "Self-supervised image representation learning with geometric set consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19270–19280.

[60] J. Hou, S. Xie, B. Graham, A. Dai, and M. Nießner, "Pri3D: Can 3D priors help 2D representation learning?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5673–5682.

[61] C.-K. Yang, M.-H. Chen, Y.-Y. Chuang, and Y.-Y. Lin, "2D–3D interlaced transformer for point cloud segmentation with scene-level supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 977–987.

[62] W. Xie, H. Jiang, S. Gu, and J. Xie, "Implicit obstacle map-driven indoor navigation model for robust obstacle avoidance," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 6785–6793.

[63] H. Du, L. Li, Z. Huang, and X. Yu, "Object-goal visual navigation via effective exploration of relations among historical navigation states," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2563–2573.

[64] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[65] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[66] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "RedNet: Residual encoder–decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*.

[67] S. Huang and S. Ontañón, "A closer look at invalid action masking in policy gradient algorithms," 2020, *arXiv:2006.14171*.

[68] B. Chen, P. Zhong, Y. Cui, S. Lu, Y. Liang, and Y. Sheng, "EMExplorer: An episodic memory enhanced autonomous exploration strategy with Voronoi domain conversion and invalid action masking," *Complex Intell. Syst.*, vol. 9, no. 6, pp. 7365–7379, Dec. 2023.

[69] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[70] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[71] J. Zhang, C. Zhu, L. Zheng, and K. Xu, "Fusion-aware point convolution for online semantic 3D scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4533–4542.

[72] S. K. Ramakrishnan, T. Nagarajan, Z. Al-Halah, and K. Grauman, "Environment predictive coding for visual navigation," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–7.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[74] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[75] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts," *Proc. Nat. Acad. Sci. USA*, vol. 93, no. 4, pp. 1591–1595, Feb. 1996.

[76] M. Savva et al., "Habitat: A platform for embodied AI research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9339–9347.

[77] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, Jul. 2020, pp. 1597–1607.
[78] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[79] E. Wijmans et al., "DD-PPO: Learning near-perfect PointGoal navigators from 2.5 billion frames," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=H1gX8C4YPr
[80] S. Zhang et al., "HOZ++: Versatile hierarchical object-to-zone graph for object navigation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5958–5975, Jul. 2025.
[81] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proc. Int. Symp. Comput. Intell. Robot. Automat. (CIRA)*, Jul. 1997, pp. 146–151.
[82] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological SLAM for visual navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12875–12884.
[83] B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging large language models for visual target navigation," 2023, *arXiv:2304.05501*.
[84] K. Zhou et al., "ESC: Exploration with soft commonsense constraints for zero-shot object navigation," 2023, *arXiv:2301.13166*.
[85] W. Cai et al., "Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill," 2023, *arXiv:2309.10309*.
[86] S. Zhang, X. Yu, X. Song, X. Wang, and S. Jiang, "Imagine before go: Self-supervised generative map for object goal navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16414–16425.
[87] P. Wu et al., "VoroNav: Voronoi-based zero-shot object navigation with large language model," 2024, *arXiv:2401.02695*.
[88] X. Zhao, W. Cai, L. Tang, and T. Wang, "ImagineNav: Prompting vision-language models as embodied navigator through scene imagination," 2024, *arXiv:2410.09874*.
[89] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[90] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[91] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 27730–27744.
[92] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=H1lJJnR5Ym

Bolei Chen is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Central South University, Changsha, China.

He has published over ten articles in journals and conferences, including IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, ACM TITS, ACM TOMM, IEEE ROBOTICS AND AUTOMATION LETTERS, ACM MM, IEEE ICRA, and IEEE IROS. His research interests include embodied AI, computer vision, multimedia, and reinforcement learning.

Jiaxu Kang received the B.S. degree from the School of Computer Science and Engineering, Central South University, Changsha, China, in 2023, where he is currently pursuing the M.S. degree.

His research interests include embodied AI and robotic navigation.

Haonan Yang received the B.S. degree from the School of Computer Science and Engineering, Central South University, Changsha, China, in 2023, where he is currently pursuing the M.S. degree.

His research interests include embodied AI and robotic navigation.

Ping Zhong (Member, IEEE) received the B.S. degree from the National University of Defense Technology, Changsha, China, in 2004, and the Ph.D. degree in communication engineering from Xiamen University, Xiamen, China, in 2011.

From 2012 to 2013, she was a Post-Doctoral Fellow with the Computer Laboratory, University of Cambridge, Cambridge, U.K. She is currently an Associate Professor with the Department of Computer Science and Technology, Central South University, Changsha. Her research interests include intelligent unmanned systems, machine learning, and the Internet of Things.

Dr. Zhong is a member of ACM, CCF, and IEICE.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in automation from Harbin Institute of Technology, Harbin, China, in 2015, and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, Bristol, U.K., in 2018.

He was a Research Associate with The Hong Kong University of Science and Technology, Hong Kong, from 2018 to 2020, and a Post-Doctoral Scholar Employee with the University of California at San Diego, San Diego, CA, USA, from 2020 to 2021. He began his faculty career as a Full Research Professor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China, in 2021, where he was promoted to a Full Professor in 2022 and attained tenure in 2024 with the College of Electronics and Information Engineering and Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis.

Dr. Fan was honored by being included in the Stanford University List of Top 2% Scientists Worldwide from 2022 to 2024, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, acknowledged as one of Xiaomi Young Talents in 2023, and received Shanghai Science and Technology 35 Under 35 Honor in 2024 as its youngest recipient.

Jianxin Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from the Central South University of Technology, Changsha, China, in 1992 and 1996, respectively, and the Ph.D. degree in computer science from Central South University, Changsha, in 2001.

He is currently a Senior Professor with the School of Computer Science and Engineering, Central South University. He has published more than 300 articles in various international journals and refereed conferences. His current research interests include artificial intelligence, algorithm analysis and optimization, machine learning, and bioinformatics.

Dr. Wang is the Chair of ACM SIGBio China.