Learning Navigational Visual Representations with Semantic Map Supervision

Yicong Hong[12,†] Yang Zhou[1] Ruiyi Zhang[1]
Franck Dernoncourt[1] Trung Bui[1] Stephen Gould[2] Hao Tan[1]
[1]Adobe Research [2]The Australian National University
mr.yiconghong@gmail.com
stephen.gould@anu.edu.au, {yazhou,ruizhang,dernonco,bui,hatan}@adobe.com
Project URL: https://github.com/YicongHong/Ego2Map-NaViT

Abstract

Being able to perceive the semantics and the spatial structure of the environment is essential for visual navigation of a household robot. However, most existing works only employ visual backbones pre-trained either with independent images for classification or with self-supervised learning methods to adapt to the indoor navigation domain, neglecting the spatial relationships that are essential to the learning of navigation. Inspired by the behavior that humans naturally build semantically and spatially meaningful cognitive maps in their brains during navigation, in this paper, we propose a novel navigational-specific visual representation learning method by contrasting the agent's egocentric views and semantic maps (Ego^2-Map). We apply the visual transformer as the backbone encoder and train the model with data collected from the large-scale Habitat-Matterport3D environments. Ego^2-Map learning transfers the compact and rich information from a map, such as objects, structure and transition, to the agent's egocentric representations for navigation. Experiments show that agents using our learned representations on object-goal navigation outperform recent visual pre-training methods. Moreover, our representations significantly improve vision-and-language navigation in continuous environments for both high-level and low-level action spaces, achieving new state-of-the-art results of 47% SR and 41% SPL on the test server.

1. Introduction

Visual representations for navigation should capture the rich semantics and the complex spatial relationships of the observations, which helps the agent to recognize visual entities and its transition in space for effective exploration. However, previous works usually adopt visual backbones which only focus on capturing the semantics of a static image, ignoring its connection to the agent or the correspondence to other views in a continuous environment [3, 9, 40, 42, 51, 68, 98]. One common approach is to apply encoders pre-trained for object/scene classification (*e.g.* ResNet [39] on ImageNet [80]/Places365 [102]), object detection (*e.g.* RCNN [38, 78] on VisualGenome [52]/MS-COCO [57]) or semantic segmentation (*e.g.* RedNet [45] on SUN RGBD [87]), or more recently, to use CLIP [72], which is trained for aligning millions of images and texts to encode agent's RGB observations [31, 46, 84]. Despite the increasing generalization ability of the features and rising zero-shot performance on novel targets, there still exists a large visual domain gap between these features and the features suitable for navigation since they lack the expressiveness of spatial relationships. For example, the connection between time-dependent observations and the correspondence between egocentric views and the spatial structure of an environment, which are important to decision making during navigation.

We suggest that there are two main difficulties in learning navigational-specific visual representations in previous research; First, there lacked a large-scale and realistic dataset of indoor environments. Popular datasets such as Matterport3D [8] and Gibson [95] provide traversable scenes rendered from real photos, but either the number of scenes is very limited, or the quality of the 3D scan is low. Synthetic datasets such as ProcTHOR [22] contain 10K generated scenes, but they are unrealistic and simple and do not capture the full complexity of real-world environments. Second, fine-tuning visual encoders while learning to navigate is very expensive because of the long traveling horizon, especially in tasks that require extensive exploration [4, 53, 63, 86]. Moreover, due to the scene scarcity, fine-tuning visual encoders is unlikely to generalize well to the novel environments. To address these problems, a large amount of research has been dedicated to augmenting the environments, either editing the agent's observations

[†]Intern at Adobe Research.

[56, 90] or adding objects [62], building new spatial structures [58], applying web-images to replace the views of the same categories [35] or generating new scenes [47, 48, 55]. Unlike augmentation, which essentially requires the agent's policy network to adapt new visual data, another strategy is to first pre-train the visual encoder before learning to navigate. Those methods usually apply images captured in 3D scenes and perform self-supervised learning or tune the encoder with proxy tasks (*e.g.* masked region modeling, angular prediction) to mitigate the domain gap and introduce spatial awareness [15, 17, 96, 98]. Furthermore, recent works tend to apply CLIP [72] to process visual inputs, as the model can provide powerful representations for grounding semantic and spatial concepts, hence facilitating learning [1, 31, 46, 83, 84, 85]. Although these methods have demonstrated significant improvements on navigation performance, they either only use more data without addressing the learning problem, not generalizable across different tasks, or ignore the spatial correspondence between consecutive frames which the agent will capture as it moves.

Inspire by the research in cognitive science that humans naturally build virtual maps from perspective observations that are helpful to track semantics, spaces and movements during navigation [29, 91, 92]; we propose a contrastive learning method between the agent's **Ego**centric view **pairs** and top-down semantic **Maps** (Ego^2-Map) for training visual encoders appropriate for navigation. Specifically, we first sample RGBD images and create corresponding semantic maps [7] from the large-scale Habitat-Matterport3D environments (HM3D) [74, 82] which contains hundreds of photo-realistic scenes, obtaining abundant and effective visual data. Then, we encode the sampled RGBD observations and the semantic maps with two separate visual backbones, respectively, and train the entire model with our Ego^2-Map contrastive learning. Once trained, the RGBD encoder will be plugged into an agent's network to facilitate visual perception. We argue that the map contains very rich and compact visual clues such as spatial structure, accessible areas and unexplored regions, object entities and their arrangement, as well as the agent's transition in the environment that are essential to navigation. Compared to existing online mapping-based approaches [10, 14, 32, 44], Ego^2-Map learning efficiently produces visual features that imply a complete map of the space known a view priori, while the online map is only a partial map. Importantly, Ego^2-Map explores a new possibility of modeling semantics and structures from simple RGBD inputs, which its resulting features are directly applicable to non-mapping-based navigational models and effectively generalizable to different tasks.

We mainly evaluate the features learned from Ego^2-Map on the Room-to-Room Vision-and-Language Navigation in Continuous Environments (R2R-CE) [3, 51] task, which requires an agent to navigate in photo-realistic environments following human natural language instructions such as "*Leave the bedroom and enter the kitchen. Walk forward and take a left at the couch. Stop in front of the window*." Addressing R2R-CE highly relies on exploiting the correspondence between contextual clues and the agent's observations, hence it is crucial for the visual encoder to provide semantic and structural meaningful representations. We found that the proposed Ego^2-Map features significantly boost the agent's performance, obtaining +3.56% and +5.10% absolute SPL improvements over the CLIP baseline [72] under the settings of high-level and low-level action spaces, respectively, and achieves the new best results on the R2R-CE test server. Moreover, our experiments show that Ego^2-Map learning also outperforms other visual representation learning methods such as OVRL [96] on the Object-Goal Navigation task (ObjNav), suggesting the strong generalization potential of the proposed methods.

2. Related Work

Visual Navigation A great variety of scenarios have been proposed to learn visual navigation in photo-realistic environments [8, 21, 54, 82, 86, 89, 95] with different modalities of inputs, targets and action spaces [3, 4, 11, 13, 51, 66, 70, 94]. Due to the distinct nature of the tasks, diverse methods are investigated accordingly. For instance, in Object-Nav, which only provides a high-level goal and requires exploration, mapping-based methods are frequently applied [10, 32, 60, 73, 105], whereas in vision-and-language navigation (VLN) [3], large vision-language models are employed to match instructions and agent's observations [42, 61, 71, 83, 103] to perform panoramic actions [30]. Recently, there is an emerging trend of scaling up the training data to address the common data scarcity issue, either in terms of the number of environments [22, 74] or the amount of supervision [16, 76, 93]. Unlike most of the previous works, we focus on improving the visual backbone, aiming to produce robust and generalizable visual representations specialized for navigation.

Visual Representations for Embodied AI Recent years have witnessed a trend of moving away from visual encoders pre-trained for object classification in Embodied AI due to their inefficiency in representing complex real-world scenes or mapping to actions. Instead, large vision-language models such as CLIP which demonstrates strong zero-shot performance across visual domains is widely applied [20, 72], enhancing the semantic representations in robotic manipulation [85], assisting 3D trajectory modification and speed control [5], benefiting the language-conditioned view-selection problem in VLN [83, 84] as well as improving other control and navigation results [31, 46, 67]. Moreover, EmbCLIP shows that the CLIP features provide much better semantic and geometric primitives such as

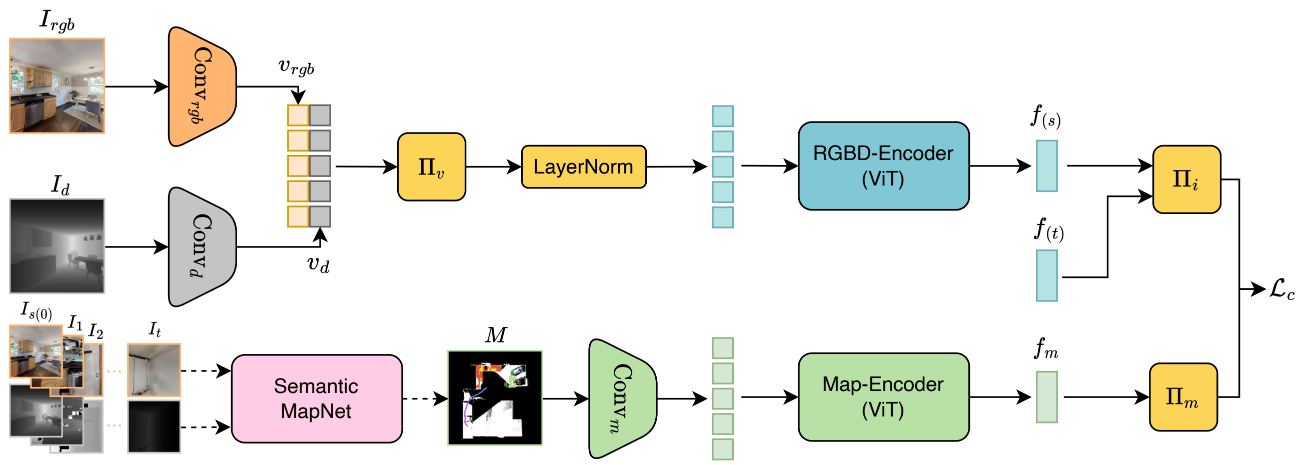


Figure 1: Network architecture. The Semantic MapNet [7] is applied to draw a semantic map using observations captured from a source to a target position. Then, two learnable visual encoders are applied to encode the egocentric views and the map, respectively, followed by non-linear headers $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_m$ to learn the Ego^2-Map contrastive objective $\mathcal{L}_c$. f_s and f_t are the paired view features of a trajectory (see §3.2). Dash arrows indicate pre-processing without gradient update.

object presence, object reachability, and free space that are valuable to Embodied AI [46]. In terms of self-supervised visual representation learning for indoor navigation, recent works that are the most relevant to ours include OVRL [96] which applies knowledge distillation based on DINO [6], EPC which predicts masked observations from a trajectory [75], and CRL which jointly learns a visual encoder with a policy network that maximizes the representation error [27]. In contrast to all previous works, this paper explores a novel idea of encoding the explicit structural and semantic information available in maps implicitly in the visual encoder thereby facilitating efficient and effective generalization across different visual navigation tasks and models.

Exploiting Visual Features A great variety of proxy tasks and auxiliary objectives have been applied to exploit information from the visual data which is beneficial to navigation. For instance, Ye *et al.* [98, 99] promote understanding of spatiotemporal relations of observations by estimating the timestep difference between two views, Qi *et al.* [69] predict room-types from object features to improve scene understanding, and Se *et al.* [81] determine the furthest reachable waypoint on a topological graph using the node observations. To extract the geometric information, Gordon *et al.* [34] use the encoded visual features to predict depth and surface normals, reconstruct RGB input, and forecast the visual features by taking certain actions. Depth prediction has also been studied by Mirowski *et al.* [64], Desai *et al.* [24] and Chattopadhyay *et al.* [12], along with learning the inverse dynamics by predicting an action taken between two sequential observations [37, 64]. Complementary to the previous methods for interpreting visual features, our Ego^2-Map learning guides a visual encoder to produce representations with expressive spatial information.

3. Contrastive Learning between Egocentric Views and Semantic Maps (Ego^2-Map)

We will first describe the overall network architecture for training the visual encoders (§3.1), followed by the details of Ego^2-Map objectives (§3.2), as well as two other widely applied spatial-aware auxiliary tasks which we investigated with our Ego^2-Map learning (§3.3). Then, we will talk about the data collection (§3.4) and network training (§3.5) processes.

3.1. Network Architecture

We build the network for Ego^2-Map learning based on a ViT-B/16 model (default initialized from CLIP [72], whose effect will be studied in §4.3). Unlike previous methods for visual navigation, which mostly employ two independent encoders to process RGB and depth images, we investigate a compact representation by feeding RGB+Depth as four-channel inputs (Figure 1). To merge the two visual modalities, we use two separate convolutional layers to encode the RGB channels I_{rgb} and the depth map I_d, respectively, followed by a token-wise concatenation and a non-linear projection $\mathbf{\Pi}_v$ to merge the resulting feature maps before feeding to the transformer layers as

$$v_{rgb} = \text{Conv}_{rgb}(I_{rgb}), \qquad v_d = \text{Conv}_d(I_d) \tag{1}$$

and

$$f = \text{ViT}\left(\mathbf{\Pi}_v \left[v_{rgb}; v_d\right]\right) \tag{2}$$

The pooled features f of the RGBD encoder are applied for learning all spatial-aware objectives $\mathcal{L}_\theta$, $\mathcal{L}_d$ and $\mathcal{L}_c$, which will be specified in the following.

3.2. Ego^2-Map Contrastive Objective

The motivation of this task is to introduce the information within a map to single-view features, offering the agent high-level semantic and spatial clues to navigate towards distant targets. Specifically, we build positive samples by coupling a views-pair (I_s and I_t from the two endpoints of a path) with a top-down semantic map M that represents the agent's transition and observations along the path, while using mismatched views-pairs and maps of different routes or environments as the negatives. Each semantic map is generated by the off-the-shelf Semantic MapNet [7], using the RGBD observations collected by an agent traveling from the source view I_s to the target view I_t via the shortest path[2]. The semantic map provides abundant information about the open space, obstacles, unexplored regions, observed objects, and the agent's action, which we consider as highly valuable and compact representations for agent navigation (see maps in Figure 2). To encode M, we apply an additional ViT-B/32 [25] with the last three transformer layers unfrozen to adapt the map images. Then, the egocentric features of the two views f_s and f_t, and the map features f_m will be passed to two MLP (multi-layer perceptrons) headers $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_m$, respectively, to compute an alignment score as:

$$c^I = \mathbf{\Pi}_i[f_s; f_t], \qquad c^M = \mathbf{\Pi}_m[f_m] \tag{3}$$

and

$$\langle c^I, c^M \rangle = \frac{c^I \cdot c^M}{\|c^I\| \|c^M\|} \tag{4}$$

Then, the InfoNCE loss [65, 101] is applied for Ego^2-Map contrastive learning. For each j-th views-map pair in a minibatch of size N, we have

$$\mathcal{L}_{c,j} = \mathcal{L}_{c,j}^{I\to M} + \mathcal{L}_{c,j}^{M\to I} \tag{5}$$

where the views to maps ($I \to M$) loss can be expressed as

$$\mathcal{L}_{c,j}^{I\to M} = -\log \frac{\exp(\langle c_j^I, c_j^M \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle c_j^I, c_k^M \rangle / \tau)} \tag{6}$$

and the map to views loss $\mathcal{L}_{c,j}^{M\to I}$ likewise. Parameter $\tau \in \mathbb{R}_+$ denotes a learnable temperature.

[2] Each resulting map is unique because the observations are determined by the unique path (actions) for connecting two views. For example, in Figure 2, only the colored regions in the map (M) are the areas seen by the agent when traveling from the source to the target positions.

3.3. Other Spatial-Aware Objectives

We investigate two additional widely applied proxy tasks with our Ego^2-Map learning. Note that we do not claim any novelty for applying these tasks but focus on studying their effect on training the representations.

Angular Offset Prediction Given two images taken from random headings at the same position, the task predicts the angular offset between them to facilitate the modeling of correspondence between views. A similar proxy task has been applied in previous works, which predicts discrete angles and uses an extra directional encoding to imply orientation [15] or regresses the agent's turning angle conditioned on language during navigation [104]. Specifically, we pass the two pooled RGBD features f_{θ_0} and f_{θ_1}, corresponding to images at two orientations of the same viewpoint, into a learnable non-linear header $\mathbf{\Pi}_\theta$ to predict the offset as $\theta^p = \mathbf{\Pi}_\theta[f_{\theta_0}; f_{\theta_1}]$. The task is learned by minimizing the mean squared error $\mathcal{L}_\theta = E[(\theta^p - \theta^*)^2]$, where θ^* is the ground-truth angular difference between $[-\pi, \pi]$, denoting either clockwise or counter-clockwise rotation whichever is closer to encourage agent's efficient rotation.

Explorable Distance Prediction To benefit the searching of explorable regions and assist obstacle avoidance, the task estimates the agent's maximal distance to forward without being blocked by any obstacle [11, 36]. Using the same notations as above, the module regresses a value from an RGBD image as $d^p = \mathbf{\Pi}_d[f]$, and is trained with supervision $\mathcal{L}_d = E[(d^p - d^*)^2]$. We cap the ground-truth distance d^* in the range of 0.5 to 5.0 meters to match the common range of the agent's depth sensors.

3.4. Data Collection

We follow the train-validation-test split of the HM3D scenes [74] and use the first 800 environments for learning. We randomly sample 252,537 viewpoint positions from the environments (see Figure 3) and render more than a million RGBD images from those positions to create 500,000 $(I_{\theta_0}, I_{\theta_1})$ views pairs as well as 500,000 (I_s, I_t, M) triplets for learning $\mathcal{L}_\theta$ and $\mathcal{L}_c$, respectively, while using all the images to learn $\mathcal{L}_d$. The positions are sampled such that all points must be located in the open space and the minimal geodesic distance between any two points is greater than 0.40 meters. Each image collected for Ego^2-Map contrastive learning is either a unique source or target view in all trajectories, and each trajectory is created by computing the shortest path between two randomly paired views within a 7 meters range. We feed the RGBD images captured from a trajectory to the Semantic MapNet [7] for drawing the corresponding semantic map. Note that the Semantic MapNet is trained on the MP3D environments [8], but we found that

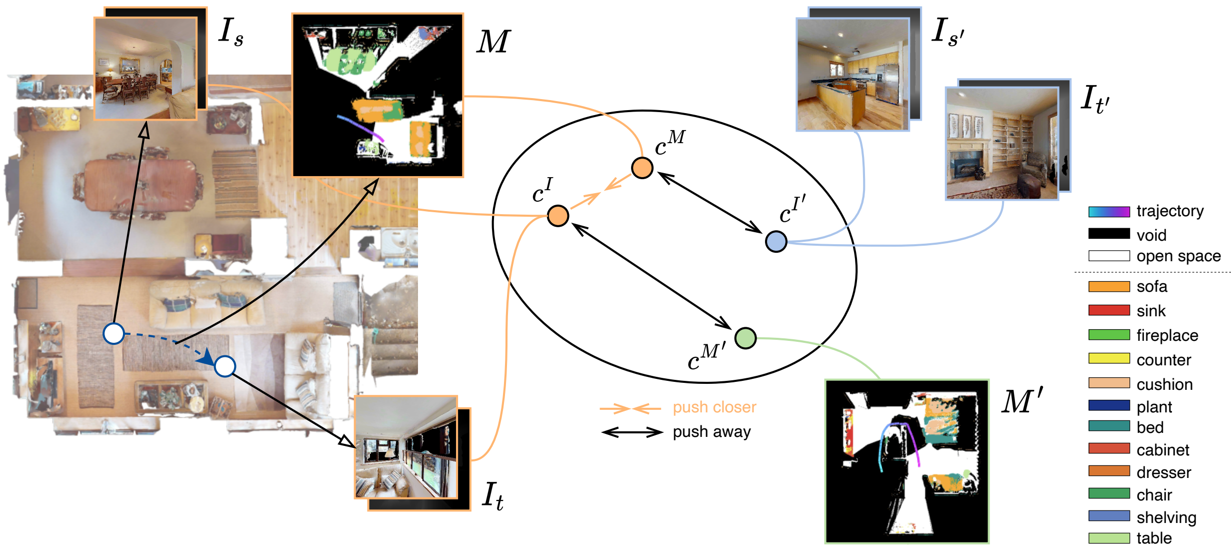
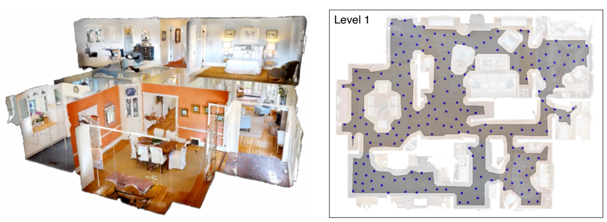


Figure 2: Illustration of Ego^2-Map contrastive learning. The method encodes the paired source and target egocentric views (I_s, I_t) into feature c^I, then learns to align it with the feature c^M of the corresponding top-down map M. Negative view pairs $(I_{s'}, I_{t'})$ and maps (M') are sampled from different trajectories or different environments, which are pushed away from the reference pair. The blue dashed line on the left part of the figure indicates the agent's trajectory, which is visualized with directional color in map M. The map is also powered with semantic information (denoted at the bottom right).



(a) 3D scan (b) Occupancy map and sampled positions

Figure 3: Illustration of an HM3D environment scan and the sampled positions (blue points) in open space (gray area).

it also generalizes well to the HM3D scenes. We refer to the *Appendix* for more sampling and dataset creation details.

3.5. Training

Implementation Details We initialize the RGBD encoder and the map encoder of our visual representation network with CLIP, a visual transformer pre-trained for aligning image-language pairs [72]. Specifically, the additional depth convolutional layer and the projection layer in the RGBD encoder are initialized randomly. All network variants in our experiments are trained with batch size 994 for 100 epochs using the AdamW optimizer [59]. After all training iterations, only the RGBD encoder will be applied as the visual backbone for navigation agents. Image augmentations are applied to all views and maps.

Optimization The overall pre-training strategy minimizes the sum over all three losses, $\mathcal{L}_c$, $\mathcal{L}_\theta$ and $\mathcal{L}_d$. All losses are equally weighted and optimized simultaneously in each iteration. After all training steps, evaluate the model with batch size 128 on 10,000 novel (I_s, I_t, M) triplets shows 92.02% views to maps $(I \rightarrow M)$ and 92.03% maps to view $(M \rightarrow I)$ alignment accuracy, indicating the correspondence between two modalities has been learned. Please see *Appendix* for more pre-training statistics and results.

4. Downstream Experiments

We evaluate our Ego^2-Map representations mainly on the R2R-CE navigation [3, 51], as addressing the task highly relies on exploiting the semantic and structural information from observations and ground to language instructions. We also show results on ObjNav [4] to compare with other recent visual representation learning approaches.

4.1. R2R-CE Setup

Our experiments consider two major setups of action spaces, either applying panorama inputs and execute high-level decisions by selecting images pointing towards navigable directions $(\mathcal{A}_{High})$[3], or using egocentric views and perform low-level controls $(\mathcal{A}_{Low})$. For $\mathcal{A}_{High}$, we adopt the candidate waypoint predictor proposed by Hong *et al.* [41], which generates navigable waypoint positions around the agent at each time step. Once a waypoint is selected, the agent will move to the position with low-level

[3]High-level VLN agents apply 36 egocentric images to represent its panoramic vision [3], hence, our Ego^2-Map features are applicable.

Models	Pre-Training Objectives			R2R-CE Val-Unseen ($\mathcal{A}_{High}$)				R2R-CE Val-Unseen ($\mathcal{A}_{Low}$)				
	Angular	Explorable	Contrastive	NE↓	nDTW↑	SR↑	SPL↑	NE↓	nDTW↑	SR↑	SPL↑	Collision↓
Baseline				4.98	57.26	48.67	42.55	7.69	49.27	25.61	23.93	17.41
1	✓			8.33	32.02	16.15	13.65	9.77	30.43	3.75	3.43	71.59
2		✓		5.83	51.25	42.58	36.34	7.94	47.15	24.74	22.98	10.44
3			✓	4.89	58.49	51.77	45.89	7.47	50.35	27.62	25.95	13.19
4		✓	✓	5.03	58.86	51.28	45.98	7.27	51.91	30.61	29.13	10.09
5	✓		✓	5.02	59.62	52.04	46.99	7.44	51.31	29.53	27.94	13.04
6	✓	✓		5.79	52.93	42.90	37.25	7.64	50.19	27.95	26.39	16.19
7	✓	✓	✓	4.94	59.67	51.77	46.11	7.25	52.01	30.40	29.03	11.25

Table 1: Ablation of Ego^2-Map learning and influence of angular and explorable objectives on the R2R-CE tasks. Checkmarks indicate applying the corresponding objectives. *Baseline* applies a pre-trained CLIP-ViT-B/16 model to encode RGB and depth separately. *Collision* measures the percentage of forward steps which the agent collided with an obstacle; we only consider this metric for $\mathcal{A}_{Low}$ since agents in $\mathcal{A}_{High}$ apply a waypoint predictor [41] to obtain navigable positions.

controls. For $\mathcal{A}_{Low}$, an agent is only allowed to do a single low-level action at each step, including Turn Left/Right 15°, Forward 0.25 meters, and Stop. In all experiments, we apply the VLN↻BERT agent [42], a vision-language transformer pre-trained for navigation, as the policy network. To use our Ego^2-Map representations, we simply replace the agent's original visual encoders (two ResNets trained on ImageNet [39] and trained in Point-Nav [94] for RGB and depth, respectively) with our RGBD encoder (§3.1). Comparing the runtime efficiency, the original encoder has around 50M parameters and the speed is about 4 GFLOPs/image, whereas our method is larger (86M parameters) but faster (20 GFLOPs/image). All visual encoders in our experiments are frozen in navigation.

R2R in continuous environment (R2R-CE) is established over the Matterport3D environments [8] based on the Habitat simulator [82]. The dataset contains 61 scans for training; 11 and 18 scans for validation and testing, respectively. R2R-CE evaluates the agent's performance using multiple metrics, including trajectory length (TL), navigation error (NE): the final distance between the agent and the target, normalized dynamic time warping score (nDTW): distance between the predicted and the ground-truth paths [43], oracle success rate (OSR): the ratio of reaching within 3 meters to the target, success rate (SR): the ratio of stopping within 3 meters to the target, and success weighted by the normalized inverse of the path length (SPL) [2].

4.2. Ablation Study

Ablation experiments in Table 1 reveal the influence of different pre-training objectives. We establish our Baseline by using a frozen CLIP-ViT-B/16 to encode RGB and depth separately, followed by a trainable fusion layer to merge the two features, which is a strong baseline that can achieve better results than the previous best encoders (see CLIP-ViT+Depth in Table 4).

Results of Model#1 indicate that employing angular offset prediction as the only task has a devastating effect on the encoder; in fact, we found that $\mathcal{L}_\theta$ only oscillates if it is minimized alone, leading to features that are not useful in downstream tasks (see details in *Appendix*). Although learning $\mathcal{L}_\theta$ with $\mathcal{L}_d$ or $\mathcal{L}_c$ can improve the performance in both $\mathcal{A}_{High}$ and $\mathcal{A}_{Low}$ (Model#5, Model#6), removing the task from Model#7 will not cause a noticeable difference (Model#4). Meanwhile, learning explorable distance prediction (Model#2, Model#5) is not effective in $\mathcal{A}_{High}$ scenario because agents with $\mathcal{A}_{High}$ apply pre-defined waypoints on open space, which means, finding explorable directions and avoid obstacle are not necessary. However, Model#2, Model#4 and Model#6 suggest that learning $\mathcal{L}_d$ with $\mathcal{L}_c$ or $\mathcal{L}_\theta$ will boost the results in $\mathcal{A}_{Low}$ and effectively reduce the collision rate during navigation.

On the other hand, by comparing the Baseline, Model#3, Model#6, and Model#7, we can see that our proposed Ego^2-Map contrastive learning has the largest impact on R2R-CE. Solely learning $\mathcal{L}_c$ improves the SR and SPL absolutely by 3.10% and 3.34% in $\mathcal{A}_{High}$, as well as 2.01% and 2.02% in $\mathcal{A}_{Low}$, while removing $\mathcal{L}_c$ from Model#7 dramatically lowers the results in both settings. Meanwhile, we found that Ego^2-Map learning can also help avoid collision, implying the useful spatial information carried by the map. In the following experiments, all three losses are applied to pre-train the visual encoders.

4.3. Discussion and Analysis

How much does the quantity of visual data matter? In Table 2, we compare the effect of sampling a different quantity of data for pre-training our visual encoder. Results show that the agent's performance in R2R-CE drops as the amount of data decreases, and reducing the number of environments has a stronger impact on the results (-1.08% SPL for 50% samples vs. -3.87% SPL for 50% envs with $\mathcal{A}_{High}$), suggesting the importance of having abundant scenes and structures in learning visual represen-

Pre-train Data	R2R-CE Val-Unseen ($\mathcal{A}_{High}$)				R2R-CE Val-Unseen ($\mathcal{A}_{Low}$)			
	NE↓	nDTW↑	SR↑	SPL↑	NE↓	nDTW↑	SR↑	SPL↑
100% data	4.94	59.67	51.77	46.11	7.25	52.01	30.40	29.03
50% samples	4.98	57.72	50.46	45.03	7.43	49.97	27.57	26.22
30% samples	5.28	56.11	48.61	42.58	7.83	48.60	25.83	24.63
10% samples	5.60	54.89	44.48	39.08	7.86	46.83	22.95	21.54
50% envs	5.32	55.53	47.80	42.24	7.57	50.35	26.54	25.15
30% envs	5.56	53.81	44.48	38.69	7.61	48.64	26.05	24.51
10% envs	5.73	53.44	41.92	36.89	7.88	47.62	23.22	22.00
MP3D	5.67	52.78	40.78	35.76	7.97	46.75	21.97	20.60

Table 2: Effect of additional environments for pre-training. The percentages indicate the portion of HM3D environments/samples for training. MP3D only uses data collected from the 61 environments in downstream.

Initialization	R2R-CE Val-Unseen ($\mathcal{A}_{High}$)				R2R-CE Val-Unseen ($\mathcal{A}_{Low}$)			
	NE↓	nDTW↑	SR↑	SPL↑	NE↓	nDTW↑	SR↑	SPL↑
CLIP	4.94	59.67	51.77	46.11	7.25	52.01	30.40	29.03
Random	4.99	57.55	49.32	43.89	7.27	51.35	29.91	28.29
IN-21K	4.95	58.61	51.55	45.73	7.05	53.30	31.05	29.69
+IN-21K map	5.40	56.18	47.09	42.01	7.21	52.68	28.55	27.13

Table 3: Effect of using pre-trained encoders for RGBD and map. *Random* and *IN-21K* initialize the RGBD encoder randomly or from a ViT pre-trained on ImageNet-21K [23], while using the CLIP-ViT-B/32 to encode maps. *+IN-21K map* also uses the IN-21K pre-trained ViT to encode the semantic map.

tations. Meanwhile, we can see that applying less than 50% of samples or environments cannot yield better results than the baseline in Table 1. This is because Ego^2-Map learning requires sufficient positive and negative data pairs to support its contrastive training, as observed in previous literatures [18, 72].

How much does initialization matter? Table 3 compares the initialization of the RGBD and the map encoders in Ego^2-Map network, either from CLIP [72], from scratch or from a ViT-B/16 [25] pre-trained on ImageNet-21K (IN-21K) [23]. We find that random initialization of the RGB encoder only leads to a slight decrease in R2R-CE performance, while initializing it from pre-trained ViT can achieve comparable results to Model#7. We further investigate these results by also replacing the initialization of our map encoder from CLIP with IN-21K pre-trained ViT; although the losses of the two models saturate to the same level in pre-training, a drastic drop is observed in R2R-CE (even lower than *Random*). We hypothesize that some semantics from CLIP can be introduced to the RGBD encoder by the map encoder through contrastive learning. While previous works often use a convolutional network to process maps [14, 32, 33, 73], there remains a valuable open question of what model is suitable for encoding maps and how to train it to benefit navigation. We will leave the study of encoding maps for future work.

Methods	R2R-CE Val-Unseen ($\mathcal{A}_{High}$)				R2R-CE Val-Unseen ($\mathcal{A}_{Low}$)			
	NE↓	nDTW↑	SR↑	SPL↑	NE↓	nDTW↑	SR↑	SPL↑
CLIP-ViT+Depth	5.47	55.55	47.15	41.28	7.93	47.76	22.68	21.69
+$\mathcal{L}$	5.24	57.74	48.67	43.78	7.07	52.59	30.07	28.74

Table 4: Pre-training a different visual encoder with Ego^2-Map objectives. The model applies CLIP-ViT-B/16 and a ResNet-50 trained for Point-Nav [94] to encode RGB and depth, respectively.

Ego^2-Map	R2R-CE Val-Unseen ($\mathcal{A}_{High}$)				R2R-CE Val-Unseen ($\mathcal{A}_{Low}$)			
	NE↓	nDTW↑	SR↑	SPL↑	NE↓	nDTW↑	SR↑	SPL↑
No semantics	5.51	57.60	47.04	42.25	7.31	51.56	29.74	28.34
No space	5.24	57.42	48.94	43.52	7.66	49.20	29.31	27.75
No target	5.02	56.26	47.20	41.21	8.05	51.04	30.83	29.07

Table 5: Ablation of information in Ego^2-Map learning.

Effect of $\mathcal{L}$ on a different encoder We further evaluate our proposed Ego^2-Map objective in pre-training a different visual encoder (Table 4). Here we consider the previous state-of-the-art encoder CLIP-ViT+Depth, which applies CLIP-ViT-B/16 and a ResNet-50 depth net trained on Gibson [95] for PointNav [82] to encode RGB and depth, respectively, following by a fusion module to integrate the encoded features. Pre-training the model with $\mathcal{L}$ leads to a significant improvement, obtaining +2.50% and +7.05% higher SPL with $\mathcal{A}_{High}$ and $\mathcal{A}_{Low}$. These results suggest that Ego^2-Map learning has the potential to be generalized to other visual encoders.

Information in Ego^2-Map contrastive learning In Table 5, we ablate the information applied in the Ego^2-Map contrastive learning. Specifically, we either remove the object semantics by masking their colored segmentations with the same color, or remove explorable regions by masking open space and void with the same color (see unmasked maps in Figure 2), or remove the target view from the view pairs to create ambiguity in agent's transition. The results show that learning Ego^2-Map without the semantic clues or open space in the maps, or without a specified target greatly damages agent performance in $\mathcal{A}_{High}$, reflecting the importance of including this information. We also find that the agent with $\mathcal{A}_{Low}$ is less sensitive to method variations in pre-training; the remaining spatial or object information on a map is still beneficial to enhance the visual representation for supporting $\mathcal{A}_{Low}$ navigation, which could be the reason why angular offset and explorable distance predictions do not show a consistent benefit in Table 1.

4.4. Comparison to Previous Methods

Advantages of Visual Representations (R2R-CE) In Table 6, we compare our method of applying Ego^2-Map RGBD encoder to the CWP-VLN↻BERT agent [41] (same model as Model#7 in Table 1) with previous approaches using $\mathcal{A}_{High}$ on the R2R-CE testing split[4]. Results show that our proposed Ego^2-Map learning brings significant improvement for all metrics, achieving a 47% SR (+3%) and a 41% SPL (+4%) over the previous best [50]. In addition, Table 7 establishes a fair comparison with methods using $\mathcal{A}_{Low}$. We can see that Ego^2-Map features greatly boost the result of the base agent CWP-VLN↻BERT [41] (23% to 30% SR), and achieve a SPL comparable to recent mapping-based methods such as CM2 [33] and WS-MGMap [14]. Note that using Ego^2-Map representations does not conflict with online mapping; while we only experiment with non-mapping-based agents, we believe the features hold great potential to facilitate modeling between views and maps in mapping-based models.

Learning Methods (ObjNav) We further compare the effect of visual representation learning methods by applying our visual encoder on ObjNav [4] as in the recent approaches [46, 96] (Table 8). All methods in the table adopt a simple pipeline as in the baseline model [94], which feeds the concatenation of visual features, GPS+Compass encodings, and the encodings of previous action to a GRU [19], then, uses a fully-connected layer to predict an action from the updated agent's state. Briefly, EmbCLIP directly uses the pre-trained CLIP-ResNet50 [72] to encode the RGB inputs, EmbCLIP-ViT+Depth applies the CLIP-ViT-B/16 to encode RGB and an extra depth net [82] pre-trained on Gibson [95] to encode the depth inputs. OVRL pre-trains the ResNet encoder with self-distillation method DINO [6], in which a student network is trained to match the output of a teacher network. Moreover, OVRL applies the Omnidata Starter Dataset (OSD)[5] [28], which is much larger and more diverse than the data for our Ego^2-Map learning (only uses the HM3D [74] subset). ObjNav measures the same metrics as R2R-CE, first, we can see that an agent using Ego^2-Map features greatly improves over the baseline and the EmbCLIP. Compared to OVRL, despite the method applies OSD for pre-training and fine-tunes the network end-to-end with human demonstrations, our method obtains a better success rate (+0.4%) and much higher SPL (+3.2%). Note that, although not directly comparable, as reported in the OVRL, the SPL on ImageNav [63] without fine-tuning the visual encoder will drop drastically from 26.9% to 17.0%, whereas all our experiments keep the visual encoder frozen during navigation. These results suggest that Ego^2-Map representations are generalizable to different navigation tasks and provide more robust visual representations.

[4]R2R-CE Challenge Leaderboard: https://eval.ai/web/challenges/challenge-page/719/leaderboard/1966

[5]OSD [28] contains approximately 14.5 million images rendered from diverse 3D environments, including Replica [88], Replica+GSO [26], Hypersim [79], Taskonomy [100], BlendedMVG [97] and HM3D [74].

Methods ($\mathcal{A}_{High}$)	R2R-CE Test-Unseen				
	TL	NE↓	OSR↑	SR↑	SPL↑
Waypoint Models [49]	8.02	6.65	37	32	30
CWP-CMA [41]	11.85	6.30	49	38	33
CWP-VLN↻BERT [41]	13.31	5.89	51	42	36
Sim2Sim [50]	11.43	6.17	52	44	37
Ego^2-Map+CWP-VLN↻BERT (ours)	13.05	**5.54**	**56**	**47**	**41**

Table 6: Comparison of agent performance on R2R-CE test server. All methods use high-level action space ($\mathcal{A}_{High}$).

Methods ($\mathcal{A}_{Low}$)	R2R-CE Val-Unseen				
	TL	NE↓	OSR↑	SR↑	SPL↑
CMA+PM+DA+Aug [51]	8.64	7.37	40	32	30
SASRA [44] †	7.89	8.32	–	24	22
LAW [77]	8.89	6.83	**44**	**35**	**31**
CWP-CMA [41] ◦	8.22	7.54	–	27	25
CWP-VLN↻BERT [41] ◦	7.42	7.66	–	23	22
CM2 [33] †	11.54	7.02	42	34	28
WS-MGMap (SemMap only) [14] †	10.89	**6.80**	42	33	28
Ego^2-Map+CWP-VLN↻BERT (ours)	8.03	7.25	37	30	29

Table 7: Comparison of agents with low-level action space ($\mathcal{A}_{Low}$) in R2R-CE Val-Unseen. †: mapping-based methods. ◦: 4% of data is removed but the comparison is still valid due to the large performance gap.

Methods	Pre-Training Dataset	ObjNav MP3D Val		
		NE↓	SR↑	SPL↑
Baseline [94]	✗	6.90	8.0	1.8
EmbCLIP [46]*	WebImageText	5.26	20.9	8.3
EmbCLIP-ViT+Depth	WebImageText+Gibson	**4.90**	23.3	8.6
OVRL no pretrain [76] †	✗	–	24.2	5.9
OVRL [96] †	OSD	–	28.6	7.4
Ego^2-Map+Baseline (ours)	HM3D	5.17	**29.0**	**10.6**

Table 8: Comparison on pre-trained visual encoders for ObjNav. † indicates the visual encoder is tuned end-to-end with behavior cloning on 40k human demonstrations collected by Habitat-Web [76], while the others freeze the visual encoder during navigation and only train the agent with DD-PPO [94] on the original data. *Results obtained by re-evaluating the officially released best model checkpoint.

5. Conclusion

In this paper, we introduce a novel method of learning navigational visual representations with contrastive learning between egocentric views pairs and top-down semantic maps (Ego^2-Map). The method transfers the compact semantic and spatial information carried by a map to the egocentric representations, which greatly facilities the agent's visual perception. Experiments show that Ego^2-Map features greatly improve the downstream navigation, such as ObjNav and VLN, and demonstrate generalization potential to different visual backbones. We believe the Ego^2-Map contrastive learning proposes a new direction of visual representation learning for navigation and provides the possibility of better modeling the correspondence between views and maps, which can further benefit agent's planning and action. Note that our work also produces a potentially effective map encoder whose full capability is worth investigating in future work.

Limitations The need to build semantic maps to enable the Ego^2-Map contrastive learning is an inevitable cost of this method; compared to other self-supervised visual representation learning approaches, Ego^2-Map requires either the semantic annotations of scenes, or traversable environments and a generalizable semantic map constructor. As a result, the data collection could be much harder. However, we also witnessed an increasing number of interactive 3D scenes being built in recent years with dense semantic annotations [22, 74, 82, 88, 95], which can facilitate scaling up our Ego^2-Map learning in the future.

References

[1] Dong An, Zun Wang, Yangguang Li, Yi Wang, Yicong Hong, Yan Huang, Liang Wang, and Jing Shao. 1st place solutions for rxr-habitat vision-and-language navigation competition (cvpr 2022). *arXiv preprint arXiv:2206.11610*, 2022. 2

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 6

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1, 2, 5

[4] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 1, 2, 5, 8

[5] Arthur Bucker, Luis Figueredo, Sami Haddadin, Ashish Kapoor, Shuang Ma, and Rogerio Bonatti. Latte: Language trajectory transformer. *arXiv preprint arXiv:2208.02918*, 2022. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3, 8

[7] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 964–972, 2021. 2, 3, 4

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017. 1, 2, 4, 6

[9] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations*, 2019. 1

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

[11] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12875–12884, 2020. 2, 4

[12] Prithvijit Chattopadhyay, Judy Hoffman, Roozbeh Mottaghi, and Aniruddha Kembhavi. Robustnav: Towards benchmarking robustness in embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15700, 2021. 3

[13] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision*, pages 17–36. Springer, 2020. 2

[14] Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. In *Annual Conference on Neural Information Processing Systems*, 2022. 2, 7, 8

[15] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:5834–5847, 2021. 2, 4

[16] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. *arXiv preprint arXiv:2208.11781*, 2022. 2

[17] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16537–16547, 2022. 2

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 7

[19] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014. 8

[20] Yuchen Cui, Scott Niekum, Abhinav Gupta, Vikash Kumar, and Aravind Rajeswaran. Can foundation models per-

form zero-shot task specification for robot manipulation? In *Learning for Dynamics and Control Conference*, pages 893–905. PMLR, 2022. 2

[21] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3164–3174, 2020. 2

[22] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. Procthor: Large-scale embodied ai using procedural generation. *arXiv preprint arXiv:2206.06994*, 2022. 1, 2, 9

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[24] Saurabh Satish Desai and Stefan Lee. Auxiliary tasks for efficient learning of point-goal navigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 717–725, 2021. 3

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4, 7

[26] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *arXiv preprint arXiv:2204.11918*, 2022. 8

[27] Yilun Du, Chuang Gan, and Phillip Isola. Curious representation learning for embodied intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10408–10417, 2021. 3

[28] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 8

[29] Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513, 2017. 2

[30] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[31] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 2022. 1, 2

[32] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. In *International Conference on Learning Representations*, 2021. 2, 7

[33] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15460–15470, 2022. 7, 8

[34] Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1022–1031, 2019. 3

[35] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643, 2021. 2

[36] Meera Hahn, Devendra Singh Chaplot, Shubham Tulsiani, Mustafa Mukadam, James M Rehg, and Abhinav Gupta. No rl, no simulation: Learning to navigate without navigating. *Advances in Neural Information Processing Systems*, 34:26661–26673, 2021. 4

[37] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations*, 2020. 3

[38] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6

[40] Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, and Stephen Gould. Language and visual entity relationship graph for agent navigation. *Advances in Neural Information Processing Systems*, 33:7685–7696, 2020. 1

[41] Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15439–15449, 2022. 5, 6, 7, 8

[42] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, June 2021. 1, 2, 6

[43] Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*, 2019. 6

[44] Muhammad Zubair Irshad, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Sasra: Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. *arXiv preprint arXiv:2108.11945*, 2021. 2, 8

[45] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 1

[46] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022. 1, 2, 3, 8

[47] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. *arXiv preprint arXiv:2204.02960*, 2022. 2

[48] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14738–14748, 2021. 2

[49] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15162–15171, 2021. 8

[50] Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 588–603. Springer, 2022. 8

[51] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, 2020. 1, 2, 5, 8

[52] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1

[53] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 1

[54] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *Conference on Robot Learning*, pages 455–465. PMLR, 2022. 2

[55] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *arXiv preprint arXiv:2305.19195*, 2023. 2

[56] Jialu Li, Hao Tan, and Mohit Bansal. Envedit: Environment editing for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15407–15417, 2022. 2

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[58] Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, Zongyuan Ge, and Yi-Dong Shen. Vision-language navigation with random environmental mixup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1644–1654, 2021. 2

[59] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[60] Haokuan Luo, Albert Yue, Zhang-Wei Hong, and Pulkit Agrawal. Stubborn: A strong baseline for indoor object navigation. *arXiv preprint arXiv:2203.07359*, 2022. 2

[61] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*, pages 259–274. Springer, 2020. 2

[62] Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. Thda: Treasure hunt data augmentation for semantic navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15374–15383, 2021. 2

[63] Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-augmented reinforcement learning for image-goal navigation. *arXiv preprint arXiv:2101.05181*, 2021. 1, 8

[64] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016. 3

[65] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[66] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022. 2

[67] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. *arXiv preprint arXiv:2203.03580*, 2022. 2

[68] Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15942–15952, 2021. 1

[69] Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1655–1664, 2021. 3

[70] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 2

[71] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: History-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15418–15427, 2022. 2

[72] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5, 7, 8

[73] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900, 2022. 2, 7

[74] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 4, 8, 9

[75] Santhosh K Ramakrishnan, Tushar Nagarajan, Ziad Al-Halah, and Kristen Grauman. Environment predictive coding for embodied agents. *arXiv preprint arXiv:2102.02337*, 2021. 3

[76] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5173–5183, 2022. 2, 8

[77] Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4018–4028, 2021. 8

[78] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[79] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10912–10922, 2021. 8

[80] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[81] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. In *International Conference on Learning Representations*, 2018. 3

[82] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 2, 6, 7, 8, 9

[83] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. *arXiv preprint arXiv:2207.04429*, 2022. 2

[84] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2021. 1, 2

[85] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022. 2

[86] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 1, 2

[87] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1

[88] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 8, 9

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

[90] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of NAACL-HLT*, pages 2610–2621, 2019. 2

[91] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948. 2

[92] Ranxiao Frances Wang and Elizabeth S Spelke. Human spatial representation: Insights from animals. *Trends in cognitive sciences*, 6(9):376–382, 2002. 2

[93] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. 2023. 2

[94] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 2, 6, 7, 8

[95] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 1, 2, 7, 8, 9

[96] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. *arXiv preprint arXiv:2204.13226*, 2022. 2, 3, 8

[97] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 8

[98] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectnav. *arXiv preprint arXiv:2104.04112*, 2021. 1, 2, 3

[99] Joel Ye, Dhruv Batra, Erik Wijmans, and Abhishek Das. Auxiliary tasks speed up learning point goal navigation. In *Conference on Robot Learning*, pages 498–516. PMLR, 2021. 3

[100] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 8

[101] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 4

[102] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1

[103] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*, 2023. 2

[104] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2020. 4

[105] Minzhao Zhu, Binglei Zhao, and Tao Kong. Navigating to objects in unseen environments by distance prediction. *arXiv preprint arXiv:2202.03735*, 2022. 2