Room-and-Object Aware Knowledge Reasoning for Remote Embodied Referring Expression

Chen Gao[1]*, Jinyu Chen[1]*, Si Liu[1]†, Luting Wang[1], Qiong Zhang[3], Qi Wu[2]
[1]Institute of Artificial Intelligence, Beihang University
[2]The University of Adelaide [3]Xiaomi AI Lab, Xiaomi Inc

Abstract

The Remote Embodied Referring Expression (REVERIE) is a recently raised task that requires an agent to navigate to and localise a referred remote object according to a high-level language instruction. Different from related VLN tasks, the key to REVERIE is to conduct goal-oriented exploration instead of strict instruction-following, due to the lack of step-by-step navigation guidance. In this paper, we propose a novel Cross-modality Knowledge Reasoning (CKR) model to address the unique challenges of this task. The CKR, based on a transformer-architecture, learns to generate scene memory tokens and utilise these informative history clues for exploration. Particularly, a Room-and-Object Aware Attention (ROAA) mechanism is devised to explicitly perceive the room- and object-type information from both linguistic and visual observations. Moreover, through incorporating commonsense knowledge, we propose a Knowledge-enabled Entity Relationship Reasoning (KERR) module to learn the internal-external correlations among room- and object-entities for agent to make proper action at each viewpoint. Evaluation on REVERIE benchmark demonstrates the superiority of the CKR model, which significantly boosts SPL and REVERIE-success rate by 64.67% *and* 46.05%*, respectively. Code is available at:* https://github.com/alloldman/CKR.

1. Introduction

The Embodied-AI (E-AI), where embodied agents perform various egocentric perception tasks, has attracted a surge of interest within both computer vision and natural language processing communities. In recent years, numerous datasets [1, 6, 16] and simulators [5, 26, 49] have been constructed to provide 3D assets with annotations and simulate the agent respectively. These platforms support legions of tasks including Vision-Language Navigation (VLN) [1], Embodied Question Answering [6], *etc*.

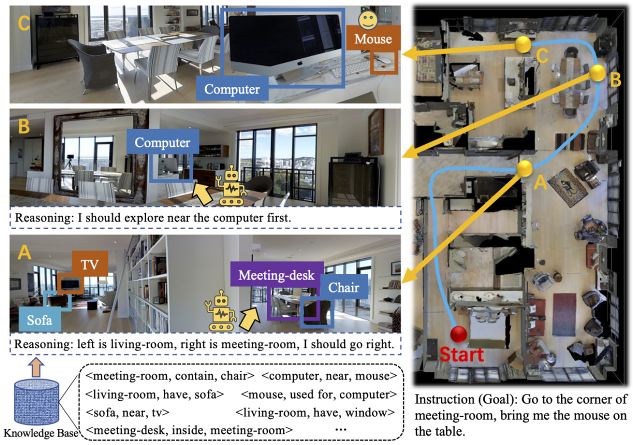
C
Mouse
Computer
B
Computer
Reasoning: I should explore near the computer first.
A
TV
Sofa
Meeting-desk
Chair
Reasoning: left is living-room, right is meeting-room, I should go right.
Knowledge Base
<meeting-room, contain, chair> <computer, near, mouse>
<living-room, have, sofa> <mouse, used for, computer>
<sofa, near, tv> <living-room, have, window>
<meeting-desk, inside, meeting-room> ···
C
B
A
Start
Instruction (Goal): Go to the corner of meeting-room, bring me the mouse on the table.

Figure 1. At viewpoint A, our agent with commonsense turns right into the 'meeting room' through perceived 'chair' and 'meeting-desk'. Then at viewpoint B, it seeks for easy-to-find related objects (*e.g*., 'computer') at first for efficient exploration, where target 'mouse' is usually around. C is the final viewpoint it arrived.

Most recently, a valuable task named Remote Embodied Visual referring Expression in Real Indoor Environments (REVERIE) [37] was proposed to further facilitate the E-AI field. The goal of REVERIE is for a robot in a photo-realistic 3D indoor environment to navigate closer to and localise a referred target object according to the given high-level instruction, which is similar to VLN task. However, simply utilising approaches in VLN is not capable of completing REVERIE task satisfactorily, which has been proved in [37] through extensive experiments.

The REVERIE contains several challenges. Firstly, essentially different from previous VLN tasks (*e.g*., R2R [1]) that provide step-by-step navigation guidance, REVERIE towards practicability only annotates high/semantic-level instructions like 'Go to the corner of meeting room, bring me the mouse on the table', as shown in Fig. 1. This is more natural and closer to the human needs from a home assistance function perspective, but it is more challenging. Therefore, instead of strict instruction-following, the agent needs to conduct *goal-oriented exploration* in an un-

*Equal contribution
†Corresponding author (liusi@buaa.edu.cn)

seen environment. Specifically, efficient exploration requires the agent to hold a long-term scene memory and extract informative memory clues for making a sequence of proper decisions. Secondly, the goal in REVERIE can be abstractly expressed as `find XXX object in XXX room', which requires the cross-modal (vision-and-language) comprehension ability for the agent to *be aware of the room/object-type* at each viewpoint, and mine their relations with the goal. Thirdly, it is non-trivial to learn the internal-external correlations among rooms and objects from limited environments and apply it to the previously-unseen environments. Thus *commonsense knowledge* is required, and an example is shown in Fig. 1. At last, due to the lack of detailed guidance, how to make the exploration more efficient, *i.e*., complete the goal in a shortest trajectory, needs to be properly considered.

In this paper, we make multi-fold innovations to address the aforementioned challenges. Firstly, we design a Cross-modality Knowledge Reasoning (CKR) model to perform the REVERIE task, where the knowledge-enabled visual and linguistic clues constitute a scene memory token. Then all the memory tokens ordered by time sequence are fed into a decoder simultaneously to predict the next action. The informative clues are effectively extracted from scene memory tokens for current decision through a learnable multi-layer attention. Secondly, we propose a Room-and-Object Aware Attention (ROAA) mechanism to explicitly recognise rooms and objects from both instruction and visual input, bridging the cross-modal semantic gap between them. Thirdly, we bring external commonsense knowledge into the REVERIE task to improve capability of capturing the complicated relationships within rooms and objects obtained under the ROAA mechanism. Specifically, we propose a Knowledge-enabled Entity Relationship Reasoning (KERR) module to incorporate prior knowledge from ConceptNet [41] for comprehensive room- and object-entity reasoning. For room-entity reasoning, we explicitly learn the room-to-room correlations to guide the action decision. For object-entity reasoning, we perform internal and external knowledge graph reasoning complementarily, where the commonsense knowledge is dynamically learned and extracted from the external graph to interact with the internal knowledge at each iteration of graph reasoning. Last but not least, we devise a Direction-Aware Loss (DAL) to penalise the actions with more angle deviation from ground truth path, and a distance-aware policy to make agent properly consider the moving distance during navigation to further improve the efficiency.

Experiments conducted on the REVERIE benchmark show our CKR model significantly boosts the SPL and REVERIE-success rate by 64.67% and 46.05% respectively on val-unseen set. Besides, extensive ablations and qualitative results verify the contribution of each component.

2. Related Work

Vision-Language Navigation. VLN that requires agent to navigate in a 3D environment following a step-by-step instruction has attracted widespread attention since it is an essential capability for a movable intelligent robot. Numerous methods [47, 48, 19, 7, 12, 10, 46, 36, 17, 20] have been proposed to address the VLN task. On the basis of Matterport3D [5], the first VLN benchmark R2R [1] was proposed, along with a multi-modal Seq2Seq baseline model. Then [11] proposed a speaker model to augment data and score candidate actions. Progress monitor [33] ensures that the notion of progress toward target is encoded by the agent. EnvDrop [44] is yet another data augmentation technique, breaking the limitation of variability of seen environments. FAST [24] allows the agent to backtrack and balances local-global signals during exploring. AuxRN [52] introduced four self-supervised tasks to improve the performance further. However, since detailed instruction is provided in VLN, the agent is required to learn how to strictly follow the step-by-step command, which is essentially different from the demand for goal-oriented exploration in REVERIE task.

Vision-Language Reasoning with External Knowledge. There has been growing interest in combining computer vision [31, 13, 9, 23, 15, 14, 29, 3, 4] and natural language processing [8, 45] techniques to perform vision-language cross-modal tasks [30, 38, 21, 22]. Especially, incorporating external knowledge for reasoning has drawn great attention recently [51, 35, 40]. Commonly used knowledge graphs (*e.g*., ConceptNet [41], DBpedia [2]) represent concepts and relationships with nodes and edges respectively. With Graph-based Neural Networks (GNN) [42, 43], knowledge can be represented in structured form, which enables interaction within visual and linguistic features. Text-KVQA [40] dataset entails the model with GGNN [28] to perform reasoning on knowledge bases. KE-GAN [35] resorts ConceptNet to calculate knowledge relation loss for generating reasonable scene parsing results. Thus the commonsense and reasoning techniques are intuitively beneficial to the REVERIE especially on unseen environments.

3. Method

3.1. Problem Setup and Overview

Problem Setup. In REVERIE [37], an agent is spawned at a random viewpoint and given an instruction $I = \{w_i\}_{i=1}^{L}$ at first, where w_i is i^{th} word and L is length. For each step t, the agent observes a panoramic view $O_t = \{o_{t,i}\}_{i=1}^{36}$, which consists of 36 divided local views. Each view $o_{t,i} = \{v_{t,i}, \theta_{t,i}, \phi_{t,i}\}$ contains an image $v_{t,i}$, a heading angle $\theta_{t,i}$ and an elevation angle $\phi_{t,i}$. Then the agent needs to take an action a_t, *i.e*., selecting one view $v_{t,k}$ from N_t navigable views $\{v_{t,i}\}_{i=0}^{N_t}$, where $v_{t,0}$ stands for the stop action. Note

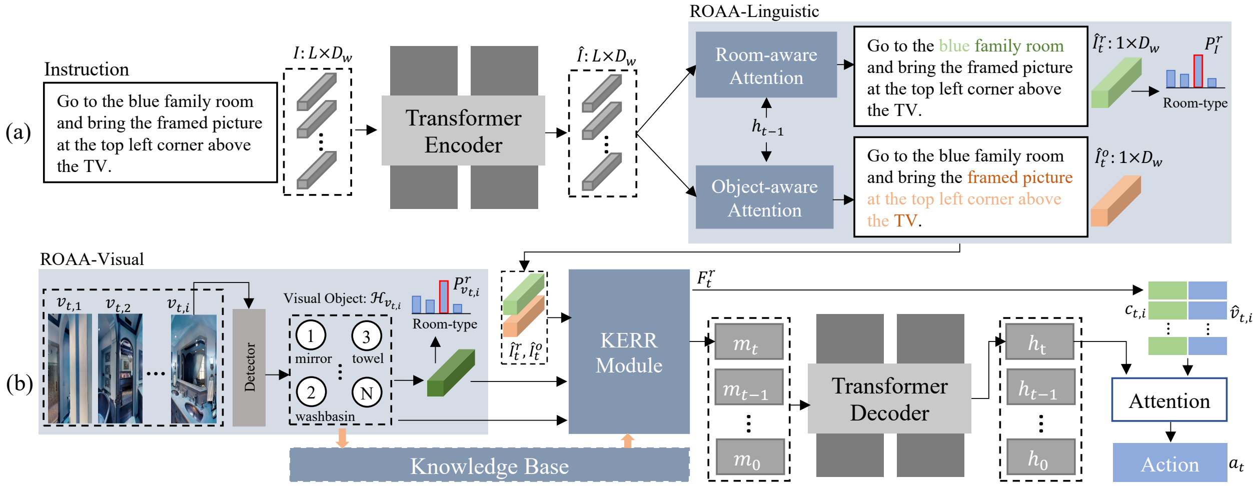
ROAA-Linguistic
Instruction
Go to the blue family room and bring the framed picture at the top left corner above the TV.
(a)
$I: L\times D_w$
Transformer Encoder
$\hat{I}: L\times D_w$
Room-aware Attention
h_{t-1}
Object-aware Attention
Go to the blue family room and bring the framed picture at the top left corner above the TV.
Go to the blue family room and bring the framed picture at the top left corner above the TV.
$\hat{I}_t^r: 1\times D_w$
P_I^r
Room-type
$\hat{I}_t^o: 1\times D_w$
ROAA-Visual
(b)
$v_{t,1}$
$v_{t,2}$
$v_{t,i}$
Detector
Visual Object: $\mathcal{H}_{v_{t,i}}$
$P^r_{v_{t,i}}$
Room-type
1 mirror
3 towel
2 washbasin
N
$\hat{I}_t^r, \hat{I}_t^o$
KERR Module
F_t^r
m_t
m_{t-1}
m_0
Transformer Decoder
h_t
h_{t-1}
h_0
$c_{t,i}$
$\hat{v}_{t,i}$
Attention
Action
a_t
Knowledge Base

Figure 2. The overall framework of CKR model. ROAA explicitly decomposes and captures the room/object feature from linguistic/visual observations, containing informative clues for REVERIE. KERR conducts room- and object-entity knowledge reasoning to make proper decisions at each viewpoint by incorporating the graph-based commonsense knowledge from ConceptNet [41]. Best viewed in color.

that $N_t \leq 36$ and is not fixed for each step. The episode ends when the agent detects an object as the target object referred in the instruction and returns its bounding box.

Overview. We propose a novel Cross-modality Knowledge Reasoning (CKR) model (see Fig. 2), which is constructed based on a transformer encoder-decoder architecture. Specifically, the encoder extracts linguistic features, and the decoder models the sequential decision process. Between the encoder and decoder, we propose a Room-and-Object Aware Attention (ROAA) mechanism to explicitly learn to decouple room- and object-related features from both linguistic and visual observations. Then the proposed Knowledge-enabled Entity Relationship Reasoning (KERR) module incorporates the external knowledge from ConceptNet [41] and takes visual clues, room-and-object cross-modal features, *etc*., to produce the scene memory token m_t at each step t. After that, the decoder takes history scene memory tokens $\{m_0, m_1, \ldots, m_{t-1}\}$ and current token m_t as inputs to produce a hidden state h_t for action prediction. Besides, the KERR module also produces a room-confidence feature $F_t^r = \{c_{t,i}\}_{i=0}^{N_t}$ for each navigable view $\{v_{t,i}\}_{k=0}^{N_t}$, where $c_{t,i}$ represents the confidence degree that moving toward $v_{t,i}$ can arrive at the target room. Further, the action a_t is predicted via an attention mechanism: $a_t = \arg\max_i(p_{t,i})$, where $p_{t,i} = softmax_i([v_{t,i}, c_{t,i}]W_a h_t^\top)$. When the agent stops, the referred object is picked from the panorama via a visual grounding model. Note that our CKR can be connected to any grounding model such as the popular MAttNet [50] and ViLBERT [32].

3.2. Room-and-Object Aware Attention

Our ROAA has two branches, focusing on linguistic (Fig. 2(a)) and visual observations (Fig. 2(b)) separately.

ROAA-Linguistic. Given a natural language instruction, we aim to focus on the room- and object-related information since the task goal can be abstracted into `'find XXX object in XXX room'`. As shown in Fig. 2(a), each word w_i of the instruction I is first initialised to a token vector by the GloVe [34] embedding. Then the transformer-encoder takes the tokens $I = \{w_i\}_{i=1}^L \in \mathbb{R}^{L\times D_w}$ along with a sequence position embedding as inputs to produce an encoded representation $\hat{I} = \{\hat{w}_i\}_{i=1}^L \in \mathbb{R}^{L\times D_w}$.

To further decompose $\hat{I}$ to room-related $\hat{I}_t^r$ and object-related $\hat{I}_t^o$ linguistic features at each step t, we adopt two language attention modules, *i.e*., room- and object-aware attention. Specifically, the two modules produce $\hat{I}_t^r, \hat{I}_t^o \in \mathbb{R}^{1\times D_w}$ by taking the decoder hidden state $h_{t-1} \in \mathbb{R}^{1\times D_h}$ as input, which is formulated as:

$$\hat{I}_t^r = \sum\nolimits_i \alpha_{t,i}\hat{w}_i, \quad \alpha_{t,i} = softmax_i(\hat{w}_i W_r h_{t-1}^\top), \tag{1}$$

$$\hat{I}_t^o = \sum\nolimits_i \beta_{t,i}\hat{w}_i, \quad \beta_{t,i} = softmax_i(\hat{w}_i W_o h_{t-1}^\top), \tag{2}$$

where $W_r, W_o \in \mathbb{R}^{D_w\times D_h}$ are learnable parameters. Then $\hat{I}_t^r$ is used to predict the probability distribution of target room-type $P_I^r = \{p_j\}_{j=1}^{N_r} = softmax(FC(\hat{I}_t^r))$ via a FC layer, where N_r is the number of room types. Note that $\hat{I}_t^r$, $\hat{I}_t^o$, and P_I^r are used in the KERR module.

ROAA-Visual. Except for understanding the goal within the instruction via ROAA-linguistic, the agent also needs to recognise the room/object from visual observations. Then the cross-modal action reasoning can be conducted based on the perceived and referred room/object during navigation.

As shown in Fig. 2(b), for i^{th} navigable view $v_{t,i}$ at step t, we adopt a Faster R-CNN [39] pre-trained on VG [27] to detect $N_{v_{t,i}}$ (≤ 100) most salient objects forming an object-set $\mathcal{H}_{v_{t,i}}$, where $|\mathcal{H}_{v_{t,i}}| = N_{v_{t,i}}$. For example, when $v_{t,i}$ is towards a bedroom, $\mathcal{H}_{v_{t,i}}$ may contain categories such as bed and wardrobe, *etc*. Note that each

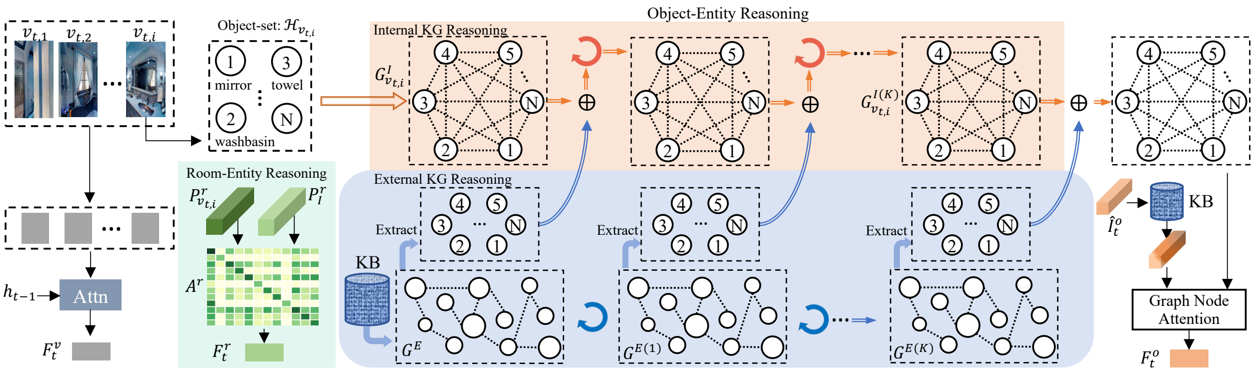

Figure 3. Illustration of the KERR module. Best viewed in color.

category is initialised as a vector by its GloVe embedding. Besides, since the type of a room mostly depends on what objects are placed in it, we take the corresponding object-set to directly predict the probability distribution of room-type through $P^r_{v_{t,i}} = \{p_j\}_{j=1}^{N_r} = softmax(FC(\mathcal{H}_{v_{t,i}}))$. Therefore, the visual room-aware features $P^r_{v_{t,i}}$ of each navigable view $v_{t,i}$ is obtained, which provides room-level informative clues for further action reasoning.

3.3. Knowledge-enabled Entity Relation Reasoning

Based on the obtained cross-modal room/object clues from both the linguistic and visual ROAA modules, we propose a KERR module aiming to model the action reasoning process. As shown in Fig. 3, the KERR produces three features: view-level F^v_t, room-level F^r_t and object-level F^o_t. The scene memory token m_t is constructed by concatenation: $m_t = [F^v_t, F^o_t, \hat{I}^r_t, \hat{I}^o_t]$, and F^r_t is directly sent to predict actions. Specifically, to obtain F^v_t, we apply a feature extractor [18] following [1] to get an image feature $\hat{v}_{t,i}$ for each navigable view $v_{t,i}$. Then we adopt h_{t-1} to generate F^v_t via an attention mechanism:

$$F^v_t = \sum\nolimits_i \eta_{t,i}\hat{v}_{t,i}, \quad \eta_{t,i} = softmax_i(\hat{v}_{t,i} W_v h^\top_{t-1}), \tag{3}$$

where W_v is a learnable parameter. F^r_t and F^o_t are produced by room- and object-entity reasoning respectively, which will be introduced in the following.

Knowledge Base Construction. To bring commonsense into our model, we construct a knowledge base (KB) according to the ConceptNet [41]. First, during navigation, the adopted detector (pre-trained on VG [27]) can distinguish 1600 categories $\{h_i\}_{i=1}^{1600}$, which includes the categories annotated in REVERIE. Then, for each category h_i, we retrieve the top-K knowledge facts $\{f_{i,j}\}_{j=1}^{K}$ from ConceptNet, where $f_{i,j} = (h_i, r_{i,j}, t_j)$, according to their relevance $r_{i,j}$. Thus the retrieved $\{t_j\}_{j=1}^{K}$ are the most closest categories to each h_i in the perspective of semantic and spatial co-occurrence. Note that we initialise each category representation via GloVe so that $h_i, t_j \in \mathbb{R}^{1\times D_w}$. Therefore, we obtain an external knowledge base represented as a graph $G^E = (\mathcal{H}^E, \mathcal{E}^E)$, where $\mathcal{H}^E$ is the label set and $\mathcal{E}^E$ is the edge set. Note that N_E (the category number of KB) includes both 1600 categories and retrieved categories. Besides, $H^E \in \mathbb{R}^{N_E \times D_w}$ represents the node feature matrix, and $A^E \in \mathbb{R}^{N_E \times N_E}$ denotes the weighted adjacency matrix, where each element $A^E_{i,j}$ is pre-defined as $r_{i,j}$.

In additional to the external knowledge graph G^E, we further define an internal knowledge graph $G^I = (\mathcal{H}^I, \mathcal{E}^I)$ to dynamically learn the domain-specific (in-door environment) knowledge in accordance with REVERIE datasets. $H^I \in \mathbb{R}^{1600\times D_w}$ denotes the node feature matrix, which is initialised via GloVe embedding. $A^I \in \mathbb{R}^{1600\times 1600}$ denotes the weighted adjacency matrix representing the correlations among objects. Unlike the pre-defined A^E, A^I is a learnable matrix and initialised from ConceptNet.

Object-Entity Reasoning. In Fig. 3, for each view $v_{t,i}$, we take its visual object-set $\mathcal{H}_{v_{t,i}}$ as the index to sample from G^I to construct a fully connected subgraph $G^I_{v_{t,i}} = (\mathcal{H}^I_{v_{t,i}}, \mathcal{E}^I_{v_{t,i}})$. $H^I_{v_{t,i}} \in \mathbb{R}^{N_{v_{t,i}} \times D_w}$ represents its node feature matrix, and $A^I_{v_{t,i}} \in \mathbb{R}^{N_{v_{t,i}} \times N_{v_{t,i}}}$ represents the learnable adjacency matrix, which is a sub-matrix of A^I.

The object-entity reasoning is achieved iteratively through two parallel graph reasoning branches, *i.e.*, external knowledge graph reasoning and internal knowledge graph reasoning, where the external knowledge is learned and dynamically *extracted* from G^E to enhance the internal knowledge reasoning. First, the external knowledge reasoning is achieved via multi-step graph convolutions:

$$\begin{cases} H^{E(k)} = \delta(A^E H^{E(k-1)} W^{E(k)}); \\ H^{E(0)} = H^E, \end{cases} \tag{4}$$

where k denotes the k^{th} step of graph reasoning and $\delta(\cdot)$ is the activation function. $W^{E(k)}$ is a learnable parameter, and $H^{E(k)}$ is node feature matrix of $G^{E(k)}$ at k^{th} step. Next, for external knowledge extracting, we take the object-set $\mathcal{H}_{v_{t,i}}$ as an index to sample a sub node feature matrix $H^{E(k)}_{v_{t,i}} \in \mathbb{R}^{N_{v_{t,i}} \times D_w}$ from $H^{E(k)}$. Then, we add $H^{E(k)}_{v_{t,i}}$ to $H^{I(k)}_{v_{t,i}}$ and conduct internal graph reasoning, which is formulated as:

$$\begin{cases} H_{v_{t,i}}^{I(k+1)} = \delta(A_{v_{t,i}}^{I} \widetilde{H}_{v_{t,i}}^{I(k)} W^{I(k+1)}); \\ \widetilde{H}_{v_{t,i}}^{I(k)} = (H_{v_{t,i}}^{I(k)} + H_{v_{t,i}}^{E(k)})/2; \\ H_{v_{t,i}}^{I(0)} = H_{v_{t,i}}^{I}. \end{cases} \tag{5}$$

Particularly, we term $H_{v_{t,i}}^{I(K)}$ as the final node feature matrix.

To obtain the final object-level feature F_t^o, we first enhance the object-related linguistic feature $\hat{I}_t^o$ by incorporating object-level clues from the knowledge base. Specifically, we compute a relevance score $\gamma_{t,i}$ between $\hat{I}_t^o$ and each category H_i^E in external KG to fuse knowledge via attention mechanism:

$$\hat{I}_t^{o'} = \sum\nolimits_i \gamma_{t,i} H_i^E, \quad \gamma_{t,i} = softmax_i(H_i^E W_f \hat{I}_t^{o\top}). \tag{6}$$

Then, we take knowledge-enhanced $\hat{I}_t^{o'}$ to further attend to $H_{v_t}^{I(K)}$, deriving F_t^o via:

$$F_t^o = softmax(\hat{I}_t^{o'} W_o H_{v_t}^{I(K)\top}) H_{v_t}^{I(K)}. \tag{7}$$

Room-Entity Reasoning. Humans are capable of perceiving the room-to-room correlation, *e.g.*, it is not a good choice to step into a toilet when the target is kitchen, instead we may find a path to the kitchen from a dining room. Therefore, we aim to equip the agent with the same capability through learning a room-to-room correlation matrix $A^r \in \mathbb{R}^{N_r \times N_r}$, where each element $A_{i,j}^r$ represents the confidence degree that agent can arrive at j^{th} room-type from i^{th} room type. Then the confidence score is produced via:

$$s_{t,i} = P_I^r A^r P_{v_{t,i}}^{r\top}, \tag{8}$$

where P_I^r is the predicted target room-type from instruction and $P_{v_{t,i}}^r$ is the predicted room-type of $v_{t,i}$. Then $s_{t,i}$ is repeated to form the confidence feature $c_{t,i} \in \mathbb{R}^{1\times 128}$ denoting a confidence degree that agent can efficiently arrive at the target room through selecting $v_{t,i}$ as the next direction. Note that $F_t^r = \{c_{t,i}\}_{i=0}^{N_t}$.

3.4. Direction-and-Distance Aware Policy

Direction-Aware Loss (DAL). When the agent fails to follow the shortest path to the target during navigation, it is expected to keep at least moving toward its target. Prior approaches, however, implicitly assume that candidate actions (excluding ground truth) are all the same, by applying a simple cross-entropy loss on each step's action selection.

In Fig. 4(a), cross-entropy loss $L_{ce} = -\log p_{t,a_1}$ is irrelevant to p_{t,a_2} and p_{t,a_3}. Since $\theta_{t,a_2} < \theta_{t,a_3}$, choosing a_2 is more likely for agent to hit the target with a shorter path, compared with a_3. Therefore, we propose a direction-aware loss to penalise the selected action with more angle deviation to ground truth path, which is formulated as:

$$L_{dir} = \sum\nolimits_{t=1}^{T} \sum\nolimits_{a=0}^{N_t} f(\theta_{t,a}) p_{t,a}, \quad f(\theta) = (1-\cos\theta)/2, \tag{9}$$

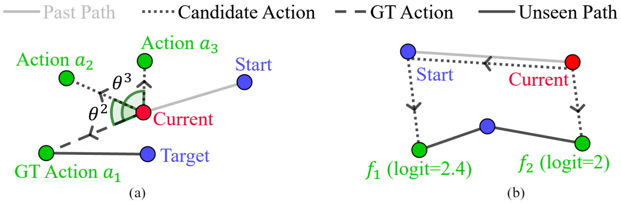

Figure 4. (a) At step t, the agent chooses an action $a_t \in \{a_0, a_1, a_2, a_3\}$ according to probability $P_{t,a}$. Note that a_0 means stop action, a_1 is the ground truth. (b) During inference, the agent chooses whether to backtrack to f_1 or proceed to f_2.

where $p_{t,a}$ is the probability of choosing action a. $\theta_{t,a} \in [0, \pi]$ is the angle between a and the GT action. Specifically, angle between a_0 and any non-stop action is $\pi/2$.

Distance-aware search. FAST [24] introduces a search policy during inference, which can achieve higher SR than greedy decoding. It records a logit for every candidate's viewpoints and proceeds to the viewpoint with largest logit at each step. However, FAST does not consider the distance between current and candidate viewpoints when making action. Therefore, we propose a distance-aware search policy based on FAST to make the agent aware of distance and avoid unnecessary long-distance backtracking (*e.g.*, prefer f_2 instead of f_1). Specifically, we leverage $logit' = logit/d^w$ instead of $logit$ during search, where d is the distance between current and candidate viewpoint, w is a hyperparameter controlling the importance of distance in making action prediction.

3.5. Training Loss

Our training objective is composed of three different parts: the imitation learning loss, the room-type classification loss and the direction-aware loss mentioned before.

Imitation Learning Loss. We leverage the student force training strategy. At step t, the model predicts the probability $p_{t,a}$ for each candidate action a, and the teacher action $\hat{a}_t$ is the shortest path from the current viewpoint to the destination. Thus the imitation learning loss is defined as:

$$L_a = \sum\nolimits_{t=1}^{T} -\log p_{t,\hat{a}_t}. \tag{10}$$

Room Classification Loss. We term $\hat{r}$ as the GT room-type of destination and $\hat{r}_{t,i}$ as the GT room-type of each view $v_{t,i}$. Thus the room classification loss is defined as:

$$L_r = \sum\nolimits_{t=1}^{T} \left(-\log p_I^{\hat{r}} + \sum\nolimits_{i=1}^{N_{v_{t,i}}} -\log p_{v_{t,i}}^{\hat{r}_{t,i}}\right). \tag{11}$$

Overall. The final objective is defined as:

$$L = \lambda_1 L_a + \lambda_2 L_r + \lambda_3 L_{dir}, \tag{12}$$

where $\lambda_i (i = 1, 2, 3)$ is the weighting factor that controls the relative importance of each term.

Methods	Val-Seen					Val-Unseen					Test (Unseen)				
	Navigation Acc.				REVERIE	Navigation Acc.				REVERIE	Navigation Acc.				REVERIE
	SR↑	OSR↑	TL↓	SPL↑	SR↑	SR↑	OSR↑	TL↓	SPL↑	SR↑	SR↑	OSR↑	TL↓	SPL↑	SR↑
Random	2.74	8.92	11.99	1.91	1.97	1.76	11.93	10.76	1.01	0.96	2.30	8.88	10.34	1.44	1.18
R2R-TF [1]	7.38	10.75	11.19	6.40	4.22	3.21	4.94	11.22	2.80	2.02	3.94	6.40	10.07	3.30	2.32
R2R-SF [1]	29.59	35.70	12.88	24.01	18.97	4.20	8.07	11.07	2.84	2.16	3.99	6.88	10.89	3.09	2.00
RCM [47]	23.33	29.44	10.70	21.82	16.23	9.29	14.23	11.98	6.97	4.89	7.84	11.68	10.60	6.67	3.67
SelfMonitor [33]	41.25	43.29	**7.54**	39.61	30.07	8.15	11.28	**9.07**	6.44	4.54	5.80	8.39	**9.23**	4.53	3.10
FAST-Short [24]	45.12	49.68	13.22	40.18	31.41	10.08	20.48	29.70	6.17	6.24	14.18	23.36	30.69	8.74	7.07
FAST-Lan-Only [24]	8.36	23.61	49.43	3.67	5.97	9.37	29.76	45.03	3.65	5.00	8.15	28.45	46.19	2.88	4.34
REVERIE [37]+FAST	50.53	55.17	16.35	45.50	31.97	14.40	28.20	45.28	7.19	7.84	19.88	**30.63**	39.05	11.61	11.28
Ours (CKR)	**57.27**	**61.91**	12.16	**53.57**	**39.07**	**19.14**	**31.44**	26.26	**11.84**	**11.45**	**22.00**	30.40	22.46	**14.25**	**11.60**

Table 1. **Main Comparisons**. Our CKR model significantly boosts the performance in terms of all the key metrics on all sets.

4. Experiments

4.1. Experimental Setup

Dataset. We conduct training and evaluation process on the REVERIE benchmark [37], which is built upon the Matterport3D simulator [1]. The dataset is split into training, validation and testing set. The training set consists of 59 scenes and 10,466 instructions over 2,353 objects. The validation-seen set (val-seen) contains 53 scenes, 1,371 instructions and 428 objects. The validation-unseen set (val-unseen) has 10 unseen scenes, 3,753 instructions and 525 objects that do not appear in the training set. The test set consists of 6,292 instructions involving 834 objects in 16 different scenes.

Evaluation Metrics. We adopt the same metrics used in [37] to evaluate models. Specifically, REVERIE Success Rate (RSR) is the key metric, where the success is achieved when the agent stops within 3 meters to the target and localises the correct object. Besides, four evaluation metrics commonly used in VLN are applied to evaluate the navigation performance: Success Rate (SR), Success rate weighted by trajectory Path Length (SPL), Oracle Success Rate (OSR) and Trajectory Length (TL), where SPL is the primary measure of navigation performance.

Implementation Details. Channel dimensions of features are set to $D_w = 300$ and $D_h = 512$. Our training process consists of two parts: 1) The CKR model. 2) The visual grounding model, where ViLBERT [32] is adopted. For the first part, we leverage the Adam optimizer [25] with weight decay 5e-4, batch size 100, learning rate 1e-4 and set $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 5$. We adopt student force strategy, which takes 10 GPU-hours to get our best model. For the second part, we fine tune ViLBERT on REVERIE by feeding instructions and RoI features at the end viewpoint. We set batch size to 60, learning rate to 2e-5 and train for 8 GPU-hours. When testing, the object with the highest matching score is predicted as the target.

4.2. Comparison with State-of-the-Art Methods

In this section, we compare our CKR model to the previous state-of-the-art methods on the REVERIE benchmark. We follow [37] to adopt previous state-of-the-art VLN approaches and apply visual grounding models on the end viewpoint during inference to complete REVERIE problem. The results are shown in Tab. 1, where we mainly compare to the RCM [47], SelfMonitor [33], FAST [24], and REVERIE [37]. Note that all methods employ MAttNet [50] fine tuned on REVERIE dataset as visual grounding approach.

In Tab. 1, the best SR and RSR performance of previous VLN+Grounding baselines on val-unseen is 10.08% and 6.24% respectively, which is quite poor and verifies the VLN methods can not achieve REVERIE well. The REVERIE [37] with a navigator-pointer method and FAST search strategy further improves the performance to 14.40% and 7.84%. Moreover, our CKR model improves all the metrics. Specifically, we improve [SR, SPL, RSR] on val-seen, val-unseen and test by $[13.33\%, 17.74\%, 22.21\%]$, $[32.92\%, 64.67\%, 46.05\%]$ and $[10.66\%, 22.74\%, 2.84\%]$. The performance gain on val-unseen is more obvious than val-seen, which benefits from the prior knowledge and reasoning process.

4.3. Ablation Experiments

We examine the contribution of each component via extensive experiments, which are shown in Tab. 2, Tab. 3 and Tab. 4. Note that 'Base-Net' in Tab. 2 denotes the transformer-based framework without any proposed modules. More ablations are in supplementary material.

ROAA Mechanism. In Tab. 2, compared with 'Base-Net', '#1' with ROAA mechanism lifts SPL from 52.29%, 7.30% to 53.78%, 8.11% on val-seen and val-unseen. The Top-5 accuracy of room-type prediction in ROAA-linguistic and -visual are shown in Tab. 4b, which also confirms its effectiveness. These results demonstrate that explicitly decoupling room/object clues benefits the navigation. Further, the visualisation results about ROAA during navigation are introduced in Sec. 4.4, which gives more insights.

Room-Entity Reasoning. In Tab. 2, comparing to '#1', '#2' confirms the effectiveness of room-entity reasoning by promoting SR from 17.92% to 18.26%. Besides, in '#3', combining object-entity reasoning can improve SR even further (19.11%). It illustrates that the two knowledge reasoning mechanisms are complementary.

Object-Entity Reasoning. In Tab. 2, comparing '#3' to

Name	ROAA	Room-Entity Reasoning	Object-Entity Reasoning	Dir-Aware	Dis-Aware	Val-Seen					Val-Unseen				
						SR↑	OSR↑	TL↓	SPL↑	RSR↑	SR↑	OSR↑	TL↓	SPL↑	RSR↑
Base-Net						56.22	62.40	13.98	52.29	37.74	16.13	29.11	43.87	7.30	9.65
#1	✓					58.47	65.43	14.03	53.78	39.21	17.92	30.13	36.26	8.11	10.96
#2	✓	✓				59.52	65.07	14.88	54.34	39.63	18.26	28.74	41.50	8.27	11.64
#3	✓	✓	✓			**65.00**	**72.94**	15.97	**58.41**	**44.13**	19.11	**40.81**	51.64	8.29	**12.92**
#4	✓	✓	✓	✓		57.27	61.49	15.77	53.14	38.72	**19.91**	33.97	37.09	10.56	12.07
#5	✓	✓	✓	✓	✓	57.28	61.91	**12.16**	53.57	39.07	19.14	31.44	**26.26**	**11.84**	11.45

Table 2. **Ablations**. The performance is gradually improved with the continuous addition of proposed modules, especially on val-unseen.

Step_1	Step_2	Val-Seen			Val-Unseen		
		SR↑	SPL↑	RSR↑	SR↑	SPL↑	RSR↑
–	–	56.57	53.02	38.17	11.62	6.77	6.32
	1	**58.40**	53.51	**40.21**	11.87	7.51	6.42
–	*3*	54.04	51.68	36.46	12.16	8.76	6.65
1	*1*	56.71	53.10	39.08	13.89	7.93	7.92
3	*3*	57.34	**54.07**	39.19	**13.94**	**9.80**	**8.72**

(a) **Object-Entity Reasoning**: Step_1 and Step_2 represents the iteration number of internal and external KG reasoning. The performance gain is obvious by conducting graph-based reasoning multi-step .

top-K	Val-Seen				Val-Unseen			
	SR↑	TL↓	SPL↑	RSR↑	SR↑	TL↓	SPL↑	RSR↑
0	57.98	11.69	54.74	37.94	12.41	15.70	8.17	8.12
5	58.05	11.96	54.17	37.98	**15.17**	16.21	**10.68**	**9.93**
10	59.66	**11.53**	**56.38**	39.04	13.92	16.10	10.10	9.11
15	**60.01**	11.86	56.03	**39.26**	13.55	**14.90**	9.76	8.87

(b) **External Knowledge Capacity**: On val-seen, the performance gain via introducing external knowledge is not obvious. On val-unseen, bring knowledge prior improves SPL by 30% when top-K=5. However, importing too much extra knowledge that also includes noise will cause performance degradation.

Table 3. **Ablations**. We mainly focus on the key metrics for each ablation setting. Note that, for simplicity, the results within this table are obtained via greedy decoding during inference instead of FAST search.

Methods	Val-Seen	Val-Unseen
[37]+*MAttNet* [50]	31.97	7.84
CKR+MAttNet [50]	37.82	11.08
CKR+ViLBERT [32]	**39.07**	**11.45**

(a)

	Room Acc.
Random	16.1%
Linguistic	97.4%
Visual	53.3%

(b)

Table 4. (a) RSR performance under different visual grounding methods. (b) Room prediction accuracy of ROAA on val-unseen.

'#2', we observe that applying the object-entity reasoning effectively improves the performance. Besides, we explore the contribution of the internal KG reasoning and external KG reasoning more deeply, which is shown in Tab. 3a. On val-unseen, applying external KG reasoning boosts SPL from 6.77% to 7.51%, and conducting multi-step (*e.g.*, 3 steps) graph reasoning can further improve SPL to 8.76%. It indicates this mechanism can effectively extract informative hierarchical clues from external KG. Moreover, by conducting multi-step internal KG reasoning to learn the domain-specific (in-door environment) knowledge, the performance is further improved to SPL=9.80%.

External Knowledge Capacity. We investigate how does the capacity of external knowledge affect the performance by varying top-K, which is shown in Tab. 3b. Note that top-K=0 represents ignoring external knowledge. On val-seen, the prior knowledge has little impact on performance. On val-unseen, the agent relies on the prior knowledge to perform better. Through setting top-K=5, SPL rises from 8.17% to 10.68%. However, continuously increasing the external knowledge (*e.g.*, top-K=10), the performance will not increase correspondingly (SPL=10.10%). It means that the external knowledge is in general-level and may contain noise that is not useful for this specific in-door domain, which also confirms the internal KG reasoning is important.

Direction-Aware Loss (DAL). In Tab. 2, compared with '#3', '#4' achieves improvement on val-unseen by importing DAL (SR from 19.11% to 19.91%, SPL from 8.29% to 10.56%). TL is also reduced from 51.64 to 37.09 meters.

Distance-aware Policy. On val-unseen in Tab. 2, '#5' significantly drops TL from 37.09 to 26.26 meters through utilising distance-aware policy. Though SR also decreases, the decline (from 19.91% to 19.14%) is smaller than that of TL, leading to a 1.28% SPL improvement.

Visual Grounding Methods. Tab. 4a shows the influence of different grounding methods. Comparing [37], our CKR equipped with the same MAttNet performs much better. By further utilising BERT-based approach [32], RSR raises to 39.07% on val-seen. But the performance gap (11.08% to 11.45%) between two methods on val-unseen is relatively small. It indicates the main improvement comes from the proposed CKR model instead of better grounding method.

4.4. Qualitative Analysis

We visualise the navigation and reasoning process, *etc.*, to illustrate more insights about our model.

Navigation Visualisation. To give a more intuitive view of how our model works during navigation, we visualise several examples in Fig. 6 and Fig. 7, where the panorama at each step (*i.e.*, viewpoint) is shown. The predicted/ground-truth room-type in each navigable view and the corresponding detected objects are shown simultaneously.

At step 1 in Fig. 6, the agent detects a hallway near a kitchen as instructed, and directs to that hallway. The agent fails to find the described phone and explores around at step 2. Then it decides to get back to the kitchen. At step 3, the agent finds an entryway which is closely connected to

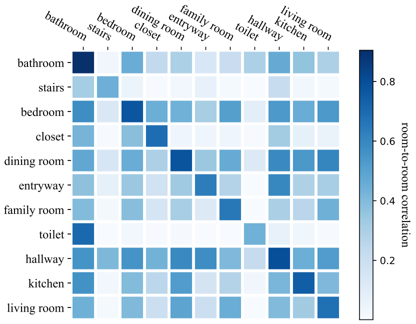

Figure 5. Learned room-to-room correlation matrix.

Category	Most relevant categories	
	ConceptNet	Learned
can	she, jar, person, containers ...	bathroom, table, trash can ...
shadow	sunset, background, cloud	bed, shoe, dog
post	column, poster, page, log	wall, bed, shoe, dog
cross	pedestrian	bathroom
bars	beam, beer	chandelier, chair
fan	pitcher mound, crowd, heater ...	chandelier, pillow, couch ...

Table 5. Illustration of the difference between general-level commonsense and domain-specific learned knowledge.

hallway in commonsense and directs to that entryway. At step 4, the agent detects a phone and decides to stop. During this process, the room-type is predicted successfully, and the agent acts reasonably for a goal-oriented exploration.

In Fig. 7(a), the agent recognises bedroom and hallway for direction. Considering goal is to go to a bathroom, and it is usually in a bedroom. Thus the agent chooses to explore bedroom first. In Fig. 7(b), the agent notices itself already in the living-room as instructed and detects related objects (*e.g*., trees). Thus it decides to stop. The agent makes proper decisions based on the observations.

Room-Entity Reasoning. We visualise the learned room-to-room correlation matrix in Fig. 5, which demonstrates the relationships among different rooms are effectively learned in the room-entity reasoning process. For example, the bedroom-to-toilet confidence is relative high since most toilets are near the bedroom.

Object-Entity Reasoning. To examine the necessity of applying internal KG reasoning to conduct domain-specific knowledge reasoning, we visualise the top-10 relevant categories in ConceptNet and the learned model (in Tab. 5). For example, the 'can' category is related to 'she' and 'person' in general, which is not useful for the REVERIE task. However, after training in REVERIE with internal KG reasoning, it is related to 'bathroom' and 'table', which shows the in-door domain knowledge is effectively learned.

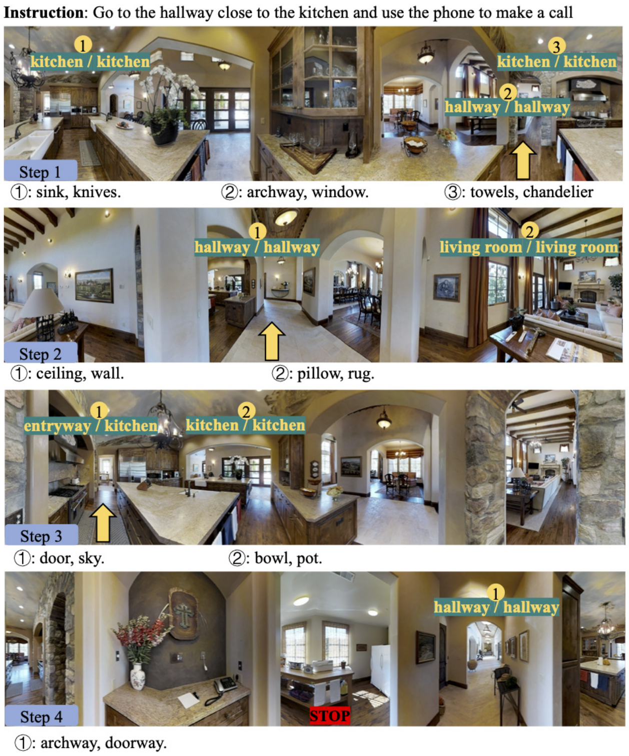

Figure 6. Visualisation of the agent behaviours on a trajectory.

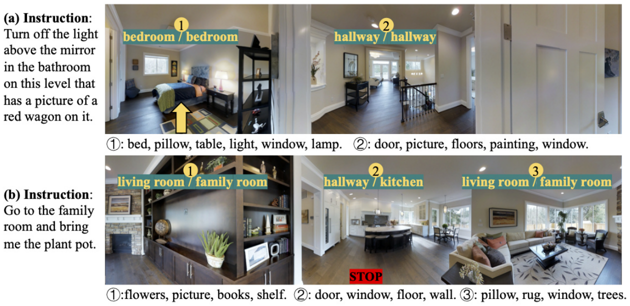
(a) Instruction: Turn off the light above the mirror in the bathroom on this level that has a picture of a red wagon on it.
bedroom / bedroom
hallway / hallway
①: bed, pillow, table, light, window, lamp.
②: door, picture, floors, painting, window.
(b) Instruction: Go to the family room and bring me the plant pot.
living room / family room
hallway / kitchen
living room / family room
STOP
①:flowers, picture, books, shelf.
②: door, window, floor, wall.
③: pillow, rug, window, trees.

Figure 7. Visualisation of the agent behaviours on two viewpoints.

5. Conclusion

In this paper, we propose a novel Cross-modality Knowledge Reasoning (CKR) model for the recently raised REVERIE task, where multi-innovations are introduced to address the unique challenges. We design a Room-and-Object Aware Attention (ROAA) mechanism to decompose room/object clues from linguistic/visual observations. We propose a Knowledge-enabled Entity Relationship Reasoning (KERR) module to conduct room/object-entity reasoning for action decision. KERR applies graph-based knowledge reasoning to capture the internal-external correlations in terms of semantic/co-occurrences among rooms/objects, where commonsense is incorporated. Extensive experiments demonstrate the superiority of our proposed methods. We believe this work will benefit and give insights to the following approaches in REVERIE and E-AI communities.

Acknowledgement This research is supported in part by National Natural Science Foundation of China under Grant 61876177, Beijing Natural Science Foundation 4202034, Fundamental Research Funds for the Central Universities, Zhejiang Lab (No. 2019KD0AB04).

References

[1] 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 *CVPR*, 2018. 1, 2, 4, 6

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. 2

[3] Jie Cao, Yibo Hu, Hongwen Zhang, Ran He, and Zhenan Sun. Learning a high fidelity pose invariant model for high-resolution face frontalization. In *NeurIPS*, 2018. 2

[4] Jie Cao, Yibo Hu, Hongwen Zhang, Ran He, and Zhenan Sun. Towards high fidelity face frontalization in the wild. *IJCV*, 2019. 2

[5] Angel X. Chang, Angela Dai, T. Funkhouser, Maciej Halber, M. Nießner, M. Savva, Shuran Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. 2017. 1, 2

[6] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *CVPR*, 2018. 1

[7] Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. Evolving graphical planner: Contextual global planning for vision-and-language navigation. In *NeurIPS*, 2020. 2

[8] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2

[9] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *CVPR*, 2020. 2

[10] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *ECCV*, 2020. 2

[11] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *NeurIPS*, 2018. 2

[12] Tsu-Jui Fu, Xin Eric Wang, Matthew F Peterson, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. Counterfactual vision-and-language navigation via adversarial path sampler. In *ECCV*, 2020. 2

[13] Chen Gao, Yunpeng Chen, Si Liu, Zhenxiong Tan, and Shuicheng Yan. Adversarialnas: Adversarial neural architecture search for gans. In *CVPR*, 2020. 2

[14] Chen Gao, Si Liu, Ran He, Shuicheng Yan, and Bo Li. Recapture as you want. In *arXiv preprint arXiv:2006.01435*, 2020. 2

[15] Chen Gao, Si Liu, Defa Zhu, Quan Liu, Jie Cao, Haoqian He, Ran He, and Shuicheng Yan. Interactgan: Learning to generate human-object interaction. In *MM*, 2020. 2

[16] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *CVPR*, 2018. 1

[17] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, 2020. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[19] Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, and Stephen Gould. Language and visual entity relationship graph for agent navigation. In *NeurIPS*, 2020. 2

[20] Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhaes, Jason Baldridge, and Eugene Ie. Transferable representation learning in vision-and-language navigation. In *ICCV*, 2019. 2

[21] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, 2020. 2

[22] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, 2020. 2

[23] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *CVPR*, 2020. 2

[24] Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *CVPR*, 2019. 2, 5, 6

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[26] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 1

[27] 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. *IJCV*, 2017. 3, 4

[28] Y. Li, Daniel Tarlow, Marc Brockschmidt, and R. Zemel. Gated graph sequence neural networks. *CoRR*, abs/1511.05493, 2016. 2

[29] Yue Liao, Si Liu, Tianrui Hui, Chen Gao, Yao Sun, Hefei Ling, and Bo Li. Gps: Group people segmentation with detailed part inference. In *ICME*, 2019. 2

[30] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *CVPR*, 2020. 2

[31] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 2

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3, 6, 7

[33] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. In *ICLR*, 2019. 2, 6

[34] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3

[35] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Kegan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *CVPR*, 2019. 2

[36] Yuankai Qi, Zizheng Pan, Shengping Zhang, and Anton van den Hengel. Object-and-action aware model for visual language navigation. In *ECCV*, 2020. 2

[37] 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 *CVPR*, 2020. 1, 2, 6, 7

[38] Guanghui Ren, Lejian Ren, Yue Liao, Si Liu, Bo Li, Jizhong Han, and Shuicheng Yan. Scene graph generation with hierarchical context. *TNNLS*, 2020. 2

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3

[40] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *ICCV*, 2019. 2

[41] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: an open multilingual graph of general knowledge. In *AAAI*, 2017. 2, 3, 4

[42] Q. Sun, Hao Peng, Jianxin Li, Senzhang Wang, Xiangyu Dong, Liangxuan Zhao, Philip S. Yu, and Lifang He. Pairwise learning for name disambiguation in large-scale heterogeneous academic networks. In *ICDM*, 2020. 2

[43] Qingyun Sun, Hao Peng, Jianxin Li, Jia Wu, Yuanxing Ning, Phillip S Yu, and Lifang He. Sugar: Subgraph neural network with reinforcement pooling and self-supervised mutual information mechanism. In *WWW*, 2021. 2

[44] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *NAACL*, 2019. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[46] Hu Wang, Qi Wu, and Chunhua Shen. Soft expert reward learning for vision-and-language navigation. In *ECCV*, 2020. 2

[47] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, 2019. 2, 6

[48] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *ECCV*, 2018. 2

[49] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *CVPR*, 2018. 1

[50] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 3, 6, 7

[51] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *ECCV*, 2020. 2

[52] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *CVPR*, 2020. 2