GeoVLN: Learning Geometry-Enhanced Visual Representation with Slot Attention for Vision-and-Language Navigation

Jingyang Huo[*], Qiang Sun[*], Boyan Jiang[*], Haitao Lin, Yanwei Fu[†]
Fudan University
jyhuo22@m.fudan.edu.cn, {18110860051, 18110240008, 19110860015, yanweifu}@fudan.edu.cn

Abstract

Most existing works solving Room-to-Room VLN problem only utilize RGB images and do not consider local context around candidate views, which lack sufficient visual cues about surrounding environment. Moreover, natural language contains complex semantic information thus its correlations with visual inputs are hard to model merely with cross attention. In this paper, we propose GeoVLN, which learns ***Geo****metry-enhanced visual representation based on slot attention for robust* ***V****isual-and-****L****anguage* ***N****avigation. The RGB images are compensated with the corresponding depth maps and normal maps predicted by Omnidata as visual inputs. Technically, we introduce a two-stage module that combine local slot attention and CLIP model to produce geometry-enhanced representation from such input. We employ V&L BERT to learn a cross-modal representation that incorporate both language and vision informations. Additionally, a novel multiway attention module is designed, encouraging different phrases of input instruction to exploit the most related features from visual input. Extensive experiments demonstrate the effectiveness of our newly designed modules and show the compelling performance of the proposed method.*

1. Introduction

With the rapid development of vision, robotics, and AI research in the past decade, asking robots to follow human instructions to complete various tasks is no longer an unattainable dream. To achieve this, one of the fundamental problems is that given a natural language instruction, let robot (agent) make its decision about the next move automatically based on past and current visual observations. This is referred as Vision-and-Language Navigation (VLN) [2]. Importantly, such navigation abilities should also work well in previously unseen environments.

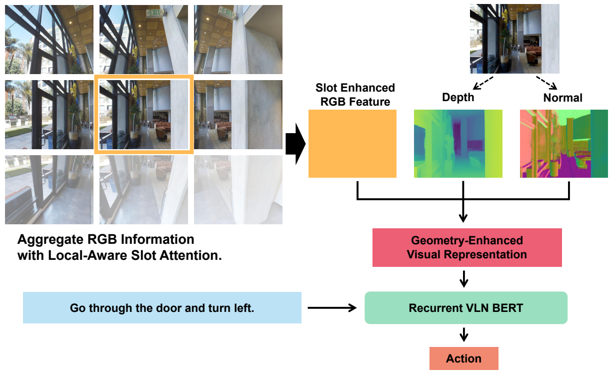
Slot Enhanced RGB Feature
Depth
Normal
Aggregate RGB Information with Local-Aware Slot Attention.
Geometry-Enhanced Visual Representation
Go through the door and turn left.
Recurrent VLN BERT
Action

Figure 1. Illustration of our learning geometry-enhanced visual representation (GeoVLN) for visual-and-language navigation. Critically, our GeoVLN utilizes the slot attention mechanism.

In the popular Room-to-Room navigation task [2], the agent is typically assumed to be equipped with a single RGB camera. At each time step, given a set of visual observations captured from different view directions and several navigation options, the goal is to choose an option as the next station. The process will be repeated until the agent reaches the end point described by the user instruction. Involving both natural language and vision information, the main challenge here is to learn a cross-modal representation that incorporates the correlations between user instruction and current surrounding environment to aid decision-making.

As solutions, early studies [2, 8, 30] resort to LSTM [13] to process temporal visual data stream. However, recent works [4, 11, 15, 18, 19, 22, 25] have taken advantage of the superior performance of the Transformer [31] and typically employ this attention-based model to facilitate representation learning with cross attention and predict actions in either recurrent [15] or one-shot [4] fashion. Despite their advantages, these approaches still have several limitations.

- 1) They only rely on RGB images which provide very

[*] Equal contributions.
[†] Corresponding authors.
Yanwei Fu is with School of Data Science, Fudan University, Shanghai Key Lab of Intelligent Information Processing, and Fudan ISTBI–ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, China.

limited 2D visual cues and lack geometry information. Thus it is hard for agent to build scene understanding about novel environments;

- 2) they process each candidate view independently without considering local spatial context, leading to inaccurate decisions;
- 3) natural language contains high-level semantic features and different phrases within an instruction may focus on various aspects visual information, *e.g*. texture, geometry. Nevertheless, we empirically find that constructing cross-modal representation with naïve attention mechanism leads to suboptimal performance.

To address these problems, we propose a novel framework, named GeoVLN, which learns **Geo**metry-enhanced visual representation based on slot attention for robust **V**isual-and-**L**anguage **N**avigation. Our framework is illustrated in Fig. 1. In particular, beyond RGB images, we also utilize the corresponding depth maps and normal maps as observations at each time step (Fig. 1), as they provide rich geometry information about environment that facilitates decision-making. Crucially, these additional mid-level cues are estimated by the recent scalable data generation framework Omnidata [7, 17] rather than sensor captured or user provided.

We design a novel two-stage slot attention [20] based module to learn geometry-enhanced visual representation from the above multimodal observations. Note that the slot attention is originally proposed to learn object-centric representation for complex scenes from single/multi-view images, but we utilize its feature learning capability and extend it to work together with multimodal observations in the VLN tasks. Particularly, we treat each candidate RGB image as a query, and choose its nearby views as keys and values to perform slot attention within a local spatial context. The key insight is that our model can implicitly learn view-to-view correspondences via slot attention, and thus encourage the candidates to pool useful features from surrounding neighbors. Additionally, we process all complementary observations, including depth maps and normal maps, through a pre-trained CLIP [26] image encoder to obtain respective latent vectors. These vectors are then concatenated with the output of slot attention module to form our final geometry-enhanced visual representation.

On the other hand, we employ BERT as language encoder to acquire global latent state and word embeddings from the input instruction. Given the respective latent embeddings for language and vision inputs, we adopt V&L BERT [11] to merge multimodal features and learn cross-modal representation for the final decision-making in a recurrent fashion following [15]. Different from previous works [21,30] that directly output probabilities of each candidate option, we present a multi-way attention module to encourage different phrases of input instruction to focus on the most informative visual observation, which boosts the performance of our network, especially in unseen environments.

To summarize, we propose the following contributions that solve the Room-to-Room VLN task with compelling accuracy and robustness:

- We extend slot attention to work on VLN task, which is combined with CLIP image encoder to learn geometry-enhanced visual representations for accurate and robust navigation.
- A novel multiway attention module encouraging different phrases of input instruction to focus on the most informative visual observation, *e.g*. texture, depth.
- We compensate RGB images with the corresponding depth maps and normal maps predicted with off-the-shelf method, improving the performance yet not involving additional training data.
- We integrate all the above technical innovations into a unified framework, named GeoVLN, and the extensive experiments validate our design choices.

2. Related Work

Vision-and-Language Navigation Exploring agents that can follow instructions to navigate in real-world scenarios is a challenging yet crucial research area for embodied artificial intelligence [6]. Promoted by the recent proposed large-scale datasets, including Matterport3D [3], Habitat [34], Gibson [28] and Room-to-Room [2], agent navigation tasks can be performed in photorealistic environments. Researchers have proposed solutions in terms of intra-modal modeling, cross-modal alignment and decision-making learning, respectively [14, 21, 32]. Early studies [2, 8, 30] use LSTM as the backbone. Benefiting from the superior performance of BERT [5], a series of recent approaches based on pre-trained vision-and-language (V&L) BERT [4, 11, 15, 18, 19, 22, 25] are introduced to the VLN task and outperform the traditional LSTM-based baselines. To name a few, PREVALENT [11] conducts V&L BERT pre-training for the image-text-action triplets and firstly derives generic representations of visual and linguistic clues applicable to VLN. Hong *et al*. [15] augments V&L BERT with a recurrent function to model the time-dependent information in the navigation process. HAMT [4] introduces a hierarchical vision transformer to capture the spatial and temporal relationships of historical observations separately, thus benefits the decision-making process. Different from prior arts [4, 15] that directly predict next action based on current multimodal conditions, we present a novel multi-way attention module inspired by MAttNet [36] to learn the correlations between the user instruction and different modalities of input observation, improving the accuracy of final decision.

Visual Representation in VLN In VLN task, directly training the whole framework end-to-end on high-resolution images is usually infeasible due to the expensive memory and computation cost. To tackle this problem, prior works [2, 24] typically utilize perceptual feature precomputed by the large pretained extractors, *e.g*. ResNet [12], and demonstrate its effectiveness. Recently, MURAL-large [16] and CLIP [26, 29] show their capability of learning expressive representations across multimodal data which facilitate several downstream tasks. We use CLIP to extract visual representation, since its feature space captures information shared by both vision and text that may encourage our model to learn semantic correspondences between them.

Additionally, some recent works [27, 35] employ slot attention [20] to learn high-level object-centric representation from single/multi-view images of a complex scene, and achieve impressive neural rendering results. Zhuang *et al*. [38] adopt vanilla slot attention module to aggregate object-centric visual information spatially. Different from [38], we newly design a novel slot-based visual representation learning module that aggregates the information from both spatial neighbors and multiple visual modalities, so that construct a better understanding of the surrounding environment for agent.

3. Method

Overview The pipeline of GeoVLN is overviewed in Fig. 2. At each time step t, our framework takes a single user instruction and a set of visual observations as input. The language input is consumed by BERT encoder [5] to obtain a global latent state $\boldsymbol{s}_0$ and a sequence of work embeddings. The visual input is composed of 36-view RGB images and the corresponding depth maps and normal maps estimated with Omnidata [7]. We design a two-stage module (Sec. 3.2) to process such multimodal observations and acquire a geometry-enhanced visual representation. Given both language and vision representations, the final action is predicted by a multiway attention based decision making module introduced in Sec. 3.3. We detail the loss functions used during training in Sec. 3.4.

3.1. Problem Setups

The Visual Language Navigation task in a discrete environment is defined on a preset connectivity graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$ denote the vertex set and edge set of the connected graph, respectively. The agent, placed at an arbitrary starting position, is asked to follow user instructions to move between vertices along the edges of the connected graph until reaching the target destination. This navigation process can be formulated concretely as follows.

Firstly, the agent is placed at a starting point in a navigable environment described by the connectivity graph. There are many different viewpoints (the number depends on specific scene) the agent can station. Then, given an instruction $\boldsymbol{I}$ containing a sequence of words, the agent is required to approach the target following the instruction. At each time step t, the agent receives observations $\boldsymbol{O}_t$ as the visual input and makes its decision about the action a_t to shift from the current state s_t to the next state s_{t+1}. It is worth noting that the observation $\boldsymbol{O}_t = \{o_t^{(i)}\}_{i=1}^{36}$ are 36 perspective projection images from different view directions rather than a complete panoramic image. These view directions have horizontal angles sampled with 30° intervals from $0° - 360°$, and pitch angles chosen from $[-30°, 0°, 30°]$. At each viewpoint, the agent is also provided with K candidate views $\boldsymbol{C}_t = \{c_t^{(i)}\}_{i=1}^{K} \subset \boldsymbol{O}_t$, corresponding to the navigable directions on the connectivity graph. The output action that the agent decided at each time step is restricted to either one of the candidate views $\boldsymbol{C}_t$ or a special “STOP” signal, denote moving to the corresponding viewpoint or decide to stop.

3.2. Two-Stage Visual Representation Learning

Visual Observations Most of the previous works only exploit RGB images as the visual observations in VLN task. However, such limited cues may involve biases about color information and lead to overfitting problem on the training environments so that hinder the generalization capability to novel scenes. To alleviate this problem, we involve other data modalities of depth maps and surface normal maps as compensation to provide geometry information that is nontrivial to be directly obtained from RGB images. Crucially, the depth maps and normal maps are estimated with the recent proposed Omnidata [7] without any additional training data.

We employ CLIP image encoder [26] pretrained on the large-scale dataset of image-text pairs to extract feature vectors (640-dimension) from all the visual observations, which are denoted as $\boldsymbol{O}_t^{rgb}$, $\boldsymbol{O}_t^{dep}$ and $\boldsymbol{O}_t^{nor}$ for RGB image, depth map and normal map respectively. We use $\{\boldsymbol{C}_t^* \mid * \in [rgb, dep, nor]\}$ to refer to the corresponding features of candidate views. Additionally, given the view angles $\{\theta, \varphi\}$ of each candidate image, we obtain an angle embedding $\boldsymbol{F}_t^{ang}$ by repeating $(\sin(\theta_i), \cos(\theta_i), \sin(\varphi_i), \cos(\varphi_i))$ 32 times as typically used in previous work [15]. We concatenate the visual features and angle features together to obtain the final candidate features:

$$\boldsymbol{F}_t^* = [\boldsymbol{C}_t^*; \boldsymbol{F}_t^{ang}], \quad * \in [rgb, dep, nor]. \tag{1}$$

And the features of the 36 views RGB observations (*i.e*. panoramic views) are:

$$\boldsymbol{P}_t^{rgb} = [\boldsymbol{O}_t^{rgb}; \boldsymbol{F}_t^{ang}], \tag{2}$$

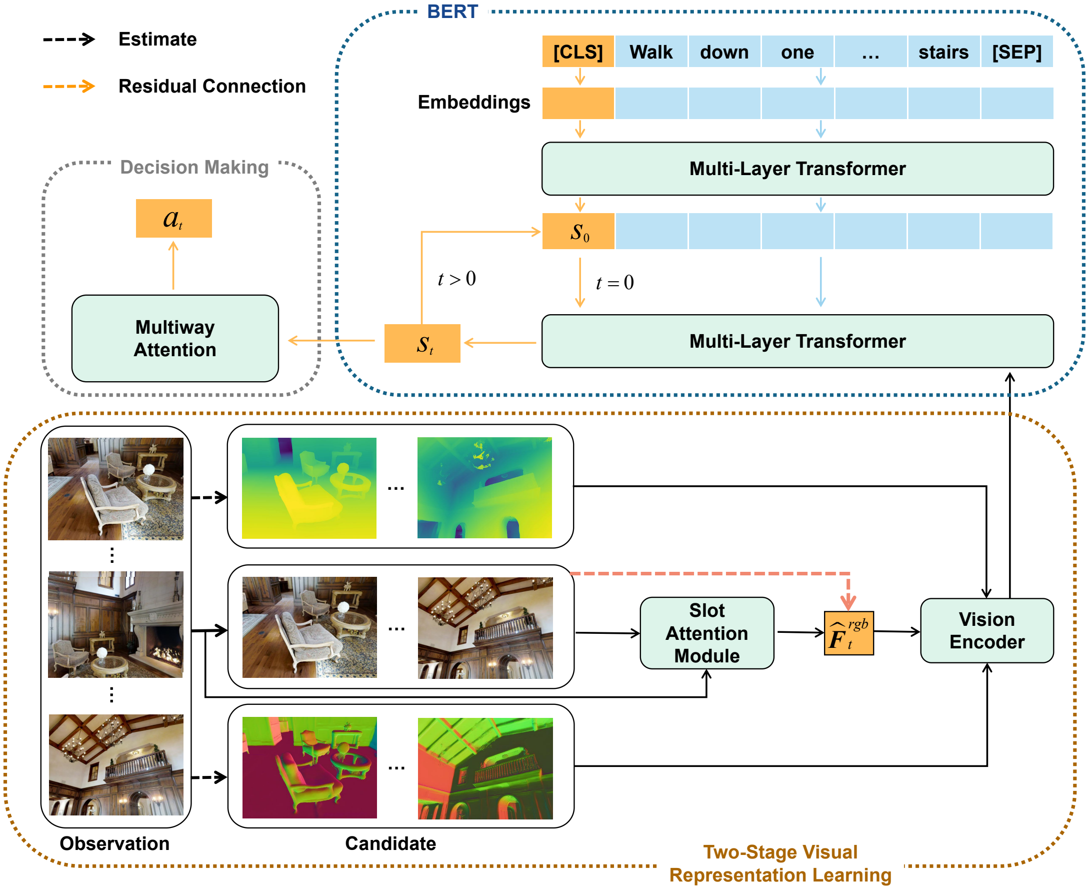

Figure 2. The overview of our GeoVLN. Particularly, at each time step, our GeoVLN takes a single user instruction and a set of visual observations as input. The language input is consumed by BERT encoder to obtain a global latent state and a sequence of work embeddings. The visual input is composed of 36-view RGB images and the corresponding depth maps and normal maps estimated with Omnidata. We have a two-stage module to process such multimodal observations and acquire a geometry-enhanced visual representation.

which are fed into the slot attention module introduced below to fuse information from local neighboring views.

Local-Aware Slot Attention Some recent works [11,15,30] only utilize the candidate views during navigation process. This brings the obstacle of understanding surround environment so that hinder navigation accuracy. For example, there are very few candidates at some viewpoints that are insufficient for making decision about next move. To mitigate this problem, we employ a slot attention module to encourage each candidate views $\boldsymbol{C}_t$ to aggregate information from the nearby observation views $\boldsymbol{O}_t$ according to the spatial proximity principle. Specifically, we initialize the slots with RGB candidate features $\boldsymbol{F}_t^{rgb}$ and treat them as queries when performing attention calculation. Additionally, the observation features $\boldsymbol{P}_t^{rgb}$ and $\boldsymbol{O}_t^{rgb}$ are used as keys and values. We apply a dropout layer to the inputs:

$$\begin{aligned} \text{slots} &= \text{Dropout}(\boldsymbol{F}_t^{rgb}), \\ Q &= \text{Dropout}(\text{LN}(\text{slots})), \\ K &= \text{Dropout}(\text{LN}(\boldsymbol{P}_t^{rgb})), \\ V &= \text{Dropout}(\text{LN}(\boldsymbol{O}_t^{rgb})), \end{aligned} \tag{3}$$

where LN denotes layer normalization.

As shown in Fig. 3, the slots are updated in a recurrent fashion following [20]. At each updating step $t = 1, \cdots, T$ ($T = 3$ in our experiments), we compute dot-product attention between keys and queries as the widely-used cross-attention, while apply Softmax operator along slot dimension to normalize the attention scores, which forces the candidate views to competitively access the information of the observations $\boldsymbol{O}_t$:

$$\text{updates} = \text{Softmax}\left(\frac{Q \cdot K^\top}{\sqrt{d_Q}}, \text{axis=slot}\right) V, \tag{4}$$

where d_Q is the dimension of Q. Then the slots are updated with a Gated Recurrent Unit (GRU) followed by a residual MLP:

$$\begin{aligned} \text{slots} &= \text{GRU}(\text{state=slots}, \text{inputs=updates}), \\ \text{slots} &= \text{slots} + \text{MLP}(\text{LN}((\text{updates})). \end{aligned} \tag{5}$$

With slot attention, the representation of each candidate view is progressively refined based on the observations

$\boldsymbol{O}_t^{rgb}$, so that the agent can capture more information from a single viewpoint to aid decision making. However, directly using all observations at one viewpoint would involve non-local information and hinder the convergence of our model. Therefore, we restrict each candidate view to focus on observations whose heading and elevation angles differ by no more than 30° from itself. We achieve this by using attention masks.

As shown in Fig. 2, we use a residual connection to add the updated slots to the candidate features $\boldsymbol{F}_t^{rgb}$. Note that we only update the visual features and keep the angle features fixed:

$$\hat{\boldsymbol{F}}_t^{rgb} = \left[\boldsymbol{C}_t^{rgb} + \text{slots}[..., : d_C]; \boldsymbol{F}_t^{ang}\right], \tag{6}$$

where d_C is the dimension of $\boldsymbol{C}_t^{rgb}$.

The output of our local-aware slot attention module is denoted as $\hat{\boldsymbol{F}}_t^{rgb}$. We then concatenate $\hat{\boldsymbol{F}}_t^{rgb}$ with the visual features from depth map and normal map, together with view angle feature, and project it into a 768-dimensional vector with a fully connected layer followed by a layer normalization.

$$\begin{aligned}\boldsymbol{F}_t &= \left[\hat{\boldsymbol{F}}_t^{rgb}[..., : d_C]; \boldsymbol{C}_t^{dep}; \boldsymbol{C}_t^{nor}; \boldsymbol{F}_t^{ang}\right] \\ \hat{\boldsymbol{F}}_t &= \text{LN}(\text{FC}(\boldsymbol{F}_t))\end{aligned} \tag{7}$$

The resulting geometry-enhanced visual representation $\hat{\boldsymbol{F}}_t$ will be used as the visual tokens of the Recurrent VLN BERT.

3.3. Multiway Attention Based Decision Making

Recurrent VLN BERT We adopt Recurrent VLN-BERT to process the instruction $\boldsymbol{I}$ and the geometry-enhanced visual representation $\hat{\boldsymbol{F}}_t$ obtained from slot attention module. At each time step, the global state vector $\boldsymbol{s}_t$ and tokens are updated with a multi-layer Transformer, which can be formulated as:

$$\boldsymbol{s}_t = \text{VLN} \circlearrowright \text{BERT}\left(\boldsymbol{s}_{t-1}, \boldsymbol{I}, \hat{\boldsymbol{F}}_t\right). \tag{8}$$

Note that $\boldsymbol{s}_t$ contains information of both vision, language as well as all the past decisions of the agent, we utilize it as the cross-modal representation to support subsequent decision-making.

Multiway Attention Different from previous works which output decisions directly from multi layer Transformer, we design a multiway attention module to compute attention scores of $\boldsymbol{s}_t$ with three modalities of visual observations: RGB, depth and normal individually, and obtain the final policy likelihood by weighted summation. We take the attention calculation with RGB features as an example. Firstly, the state representation $\boldsymbol{s}_t$ is directly projected into a 768-dimensional latent vector, while the RGB features $\hat{\boldsymbol{F}}_t^{rgb}$ are normalized by a LayerNorm (LN) operation and then projected to the same dimension through a fully connected (FC) layer:

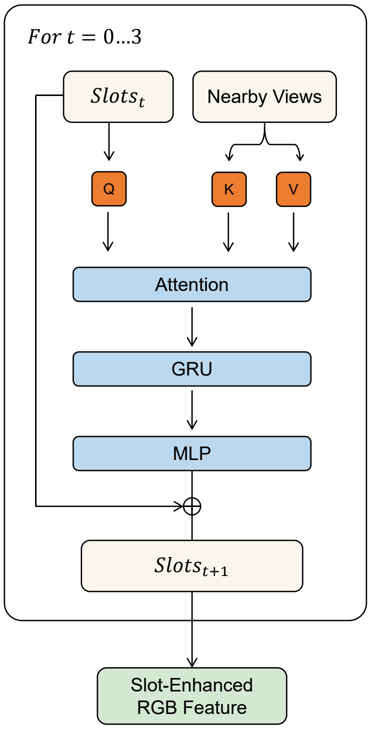

Figure 3. Detailed architecture of our local-aware slot attention module.

$$\begin{aligned}\tilde{\boldsymbol{s}}_t^{rgb} &= \boldsymbol{s}_t \boldsymbol{W}^{s,rgb}, \\ \tilde{\boldsymbol{F}}_t^{rgb} &= \text{FC}(\text{LN}(\hat{\boldsymbol{F}}_t^{rgb})).\end{aligned} \tag{9}$$

Then, the attention score can be computed as:

$$\boldsymbol{A}_t^{rgb} = \frac{\tilde{\boldsymbol{F}}_t^{rgb} \tilde{\boldsymbol{s}}_t^{rgb\top}}{\sqrt{d_h}}, \tag{10}$$

where d_h denotes the dimension of the hidden space. Similarly, the attention scores $\boldsymbol{A}_t^{dep}$ and $\boldsymbol{A}_t^{nor}$ can be obtained for the depth features and the normal features, respectively.

Matching Score At each time step, how much each modality contributes to the navigation should differ noticeably. For instance, the agent may focus more on the depth information when executing the instruction "go through the corridor" whereas the process of "picking up the spoon" will be more pertinent to the RGB and normal information. To achieve this, we apply a fully connected layer followed by a Softmax operation to compute the weights corresponding to the three modalities:

$$\left[w_t^{rgb}, w_t^{dep}, w_t^{nor}\right] = \text{Softmax}\left(\boldsymbol{s}_t \boldsymbol{W}^m + \boldsymbol{b}^m\right), \tag{11}$$

where $\boldsymbol{W}^m$ and $\boldsymbol{b}^m$ are learnable parameters.

Thus, the final matching scores of the candidate views $\boldsymbol{C}_t$ w.r.t. the state vector $\boldsymbol{s}_t$ can be written as:

$$\boldsymbol{S}_t^{total} = w_t^{rgb}\boldsymbol{A}_t^{rgb} + w_t^{dep}\boldsymbol{A}_t^{dep} + w_t^{nor}\boldsymbol{A}_t^{nor}, \qquad (12)$$

where $\boldsymbol{S}_t^{total}$ is a $(K+1)$-dimensional vector and K denotes the number of candidate views at current viewpoint. We use $\boldsymbol{S}_{t,i}^{total}(1 \leq i \leq K)$ to denote the matching score of the i-th candidate view, while $\boldsymbol{S}_{t,K+1}^{total}$ denotes the matching score of the "STOP" action. We denote the action probabilities as $\boldsymbol{p}_t$ obtained by applying the softmax function to $\boldsymbol{S}_t^{total}$. The candidate view with the highest probability is then selected as the final decision.

3.4. Loss Function

We follow the training protocol used in [15], which combines imitation learning and reinforcement learning. Specifically, our objective functions is composed of two parts. The first part is the cross-entropy loss derived from the teacher-forcing method [33]. The teacher actions are determined by the human-labeled ground-truth trajectories. Denoting the teacher action as a^*, the loss of imitation learning can be formulated as:

$$\mathcal{L}_{IL} = -\sum_t a_t^* \log\left(\boldsymbol{p}_t\right). \qquad (13)$$

Secondly, we use the A2C [23] algorithm identical to the one set in [15]. At each time step, an action is sampled according to $\boldsymbol{S}_t^{total}$ and a reward strategy is applied following the set-up. The reinforcement learning loss (Eq. (14)) is composed of three components: an actor loss to optimize strategy, a critic loss to estimate the state vector, and a regular loss to reduce action uncertainty. Additional details can be found in [15, 23].

$$\mathcal{L}_{RL} = \sum_t \mathcal{L}_{\text{actor}}^{(t)} + \mathcal{L}_{\text{critic}}^{(t)} + \lambda_{\text{reg}}\,\mathcal{L}_{\text{reg}}^{(t)} \qquad (14)$$

The overall objective function guiding our training process is

$$\mathcal{L} = \mathcal{L}_{RL} + \lambda\mathcal{L}_{IL}, \qquad (15)$$

where λ denotes the loss weight to balance both terms.

4. Experiments

4.1. Experimental setup

Dataset We use R2R [2] dataset for training and evaluation. The R2R dataset is built on 90 real-world indoor environments where the agents should traverse multiple rooms in a building to reach the destinations. And the navigation tasks are specifically described by 7189 trajectories and the corresponding instructions with the average length of 29 words. The dataset is divided into four sets including train, val seen, val unseen and test unseen sets, which mainly focus on the generalization capability of navigation in unseen environments.

Evaluation Metrics We adopt the standard metrics used in previous works for evaluation: 1) Trajectory Length (TL): the average navigational trajectory length in meters; 2) Navigation Error (NE): the distance between the final position of the agent and the target; 3) Success Rate (SR): the ratio of agents eventually stopping within 3 meters of the destination; and 4) Success Weighted by Path Length (SPL) [1] : SR weighted by the inverse of TL which measures how closely a trajectory aligns with the shortest path. A higher SPL score indicates a better balance between achieving the goal and taking the shortest path.

Implementation Details Our multimodal visual features (including RGB, depth and surface normal features) are extracted by the pretrained CLIP-Res50x4 model [26]. In the mixture of imitation learning and reinforcement learning training process, λ is set to be 0.2. For fair comparisons, we follow the training pattern of the Recurrent VLN-BERT by mixing the original training data and the augmented data with 1:1 ratio. The experiments are performed on a single GeForce GTX TITAN X GPU with AdamW optimizer. To stabilize the gradient and accelerate the convergence, we use cosine annealing scheduler with warmup and set the maximum learning rate to 10^{-5}. We train the network for 100,000 iterations with the batch size of 8 and then choose the model with the highest SPL on the validation unseen split for testing.

4.2. Main Results

The main goal of R2R VLN task is to make optimal choice at each viewpoint based on past and current information, and find the best path towards target. In this section, we provide the comparisons with previous works to show the effectiveness of the proposed GeoVLN.

Competitors. As baselines, we choose all the methods reported in [15] with the additional recent work AirBERT [10] and HAMT [4]. In addition, we extend HAMT with our newly designed modules (*i.e.* GeoVLN†) to test the effectiveness of these modules, as our contributions in GeoVLN is actually orthogonal to [4]. Specifically, we directly use the Two Stage Visual Representation Learning Module to augment the original RGB features in the HAMT, and replace the original fully connected layer with our Multiway Attention Module to make decision. Further details on how we incorporate these modules into the HAMT model can be found in the Supplementary Material.

The quantitative results are shown in Tab. 1. We mainly focus on the scores of SR and SPL in unseen environments, which provide a comprehensive evaluation of the general-

	Val Seen				Val Unseen				Test Unseen			
Agent	TL↓	NE↓	SR↑	SPL↑	TL↓	NE↓	SR↑	SPL↑	TL↓	NE↓	SR↑	SPL↑
RANDOM [2]	9.58	9.45	16	-	9.77	9.23	16	-	9.93	9.77	13	12
Human	-	-	-	-	-	-	-	-	11.85	1.61	86	76
Seq-to-Seq [2]	11.33	6.01	39	-	**8.39**	7.81	22	-	**8.13**	7.85	20	18
Speaker-Follower [8]	-	3.36	66	-	-	6.62	35	-	14.82	6.62	35	28
Self-Monitoring [9]	-	-	-	-	-	-	-	-	18.04	5.67	48	35
Reinforced Cross-Modal [32]	10.65	3.53	67	-	11.46	6.09	43	-	11.97	6.12	43	38
EnvDrop [30]	11.00	3.99	62	59	10.70	5.22	62	48	11.66	5.23	51	47
AuxRN [37]	-	3.33	70	67	-	5.28	55	50	-	5.15	55	51
PREVALENT [11]	**10.32**	3.67	69	65	10.19	4.71	58	53	10.51	5.30	54	51
PRESS [18]	10.35	3.09	71	67	10.06	4.31	59	55	10.52	4.53	57	53
AirBERT [10]	11.09	2.68	75	70	11.78	4.01	62	56	12.41	4.13	62	57
VLN ↻ BERT [15]	11.13	2.90	72	68	12.01	3.93	63	57	12.35	4.09	63	57
GeoVLN (Ours)	11.98	3.17	70	65	11.93	3.51	67	61	13.02	4.04	63	58
HAMT [4]	11.15	2.51	76	72	11.46	**2.29**	66	61	12.27	**3.93**	**65**	60
GeoVLN† (Ours)	10.68	**2.22**	**79**	**76**	11.29	3.35	**68**	**63**	12.16	3.95	**65**	**61**

Table 1. Comparison of **OUR MODEL** with the previous state-of-the-art methods on R2R dataset. † indicates the results with HAMT as the backbone. The primary metric is SPL.

	Input			Val Seen		Val Unseen	
Model	RGB	DEPTH	NORMAL	SR↑	SPL↑	SR↑	SPL↑
Baseline	✓			**69.83**	64.21	64.50	58.35
Baseline	✓	✓		66.99	63.28	63.86	58.58
Baseline	✓		✓	68.46	63.66	62.71	57.20
Baseline	✓	✓	✓	66.41	62.51	64.75	59.31
LSA	✓			67.58	63.06	64.62	59.78
LSA	✓	✓	✓	68.66	63.92	66.54	60.62
LSA + MAtt	✓			68.46	63.92	66.11	60.31
LSA + MAtt (Full)	✓	✓	✓	69.64	**64.86**	**66.75**	**61.00**

Table 2. Ablation study on multi-modal visual inputs and LSA module with VLN ↻ BERT as the backbone.

ization capability. Additionally, we also report the metrics of TL and NE.

Our results, presented in Tab. 1, demonstrate the effectiveness of our proposed models based on VLN ↻ BERT and HAMT as the backbone network. Notably, our models achieve the best performance overall, outperforming all baseline methods, with particularly impressive results in unseen environments. In comparison to VLN ↻ BERT [15] on which our framework is built, GeoVLN improves SPL and SR by 7.0% and 6.3%, respectively, on the val-unseen split. Similarly, when compared to HAMT, our proposed modules lead to significant improvements of 3.3% and 3.0% on SPL and SR, respectively. These results demonstrate the efficacy of our GeoVLN approach. While our performance is slightly inferior to VLN ↻ BERT on the val-seen split, our experimental results support our claims and contributions.

Further, our geometry-enhanced visual representation is derived from object-centric learning [20]. Unfortunately, we notice that most of previous works can only successfully work the slot attention mechanism on synthetic dataset, *e.g.* CLEVR3D. In contrast, we extend it to VLN task to encourage feature fusion between spatially neighboring views. It is capable of working under the complex real-world environments of R2R dataset.

Furthermore, to reveal the insights, we provide an ablation study with visualization results about our local slot attention in the following subsections.

4.3. Ablation Study

In this section, we provide extensive ablation studies to validate the effectiveness of novel technical designed in our GeoVLN. For fair comparisons, all the variants in our experiment follow the same training setup described in Sec. 4.1.

Our quantitative results with VLN ↻ BERT as the backbone are presented in Tab. 2. The "baseline" denotes VLN ↻ BERT trained with the RGB features extracted by CLIP as the visual representation.

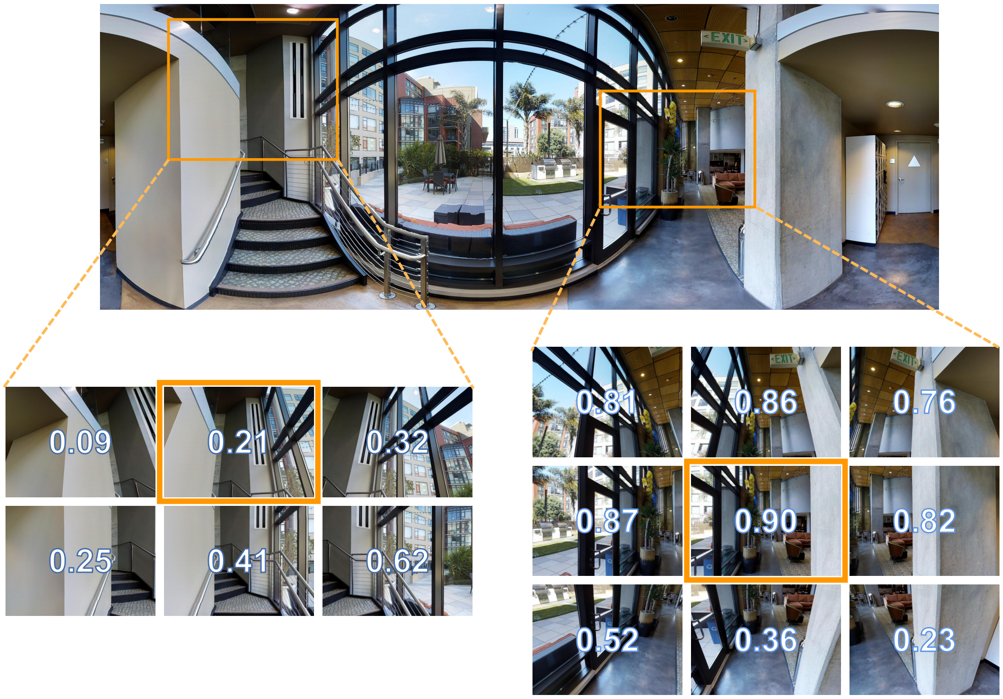

Figure 4. An visualization example of Walk up stairs. It shows the effectiveness of our local-aware slot attention module.

Firstly, we change the composition of different types of visual inputs, the results show that merely add depth map or normal map cannot offer better performance. This is possibly because the estimated depth and normal may contain errors which do not perfectly match RGB captures, so that hinder test accuracy. However, when both depth and normal inputs are given, they can benefit each other and provide geometry information that facilitates navigation. This is evidenced by improved performance in the Baseline, LSA, and LSA+MAtt models when depth and normal inputs are provided, highlighting the effectiveness of multi-modal visual inputs like depth and normal.

Next, we demonstrate the efficacy of our local-aware slot attention (LSA) and multiway attention (MAtt) by adding them to the baseline model one by one. The results show that the inclusion of LSA improves SPL by 2.5% on the val-unseen split. And by incorporating our MAtt module, our full model facilitates the identification of the most relevant visual modality for different phrases, resulting in superior performance compared to the baseline. Notably, MAtt provides valuable interpretability for decision-making processes, as demonstrated in the Supplementary material.

4.4. Visualization

To further show the effectiveness of our local-aware slot attention module, we show an visualization example in Fig. 4. The panoramic image above shows the whole room, and we choose two candidate view (in orange bounding box) for visualization below. The number on each image denotes attention score.

As shown on the left, the agent arrives at the location of stairs, and it needs information about stairs and handrail as reference to make decision of next move. So the nearby images containing stairs or handrail have higher attention score, which means that the agent successfully obtains useful features from local neighbors to aid decision-making. More visualization results are shown in Supplementary material, which illustrate the effectiveness of both our local-aware slot attention module and multiway attention module.

5. Conclusions

This paper introduces GeoVLN, which learns **Geo**metry-enhanced visual representation based on slot attention for robust **V**isual-and-**L**anguage **N**avigation. We compensate RGB captures with the estimated depth maps and normal maps as visual observations, and design a novel two-stage slot-based module to learn geometry-enhanced visual representation. Moreover, a multiway attention module is presented to facilitate decision-making. Extensive experiments on R2R dataset demonstrate the effectiveness of our newly designed modules and show the compelling performance of the proposed method.

Acknowledgement. This work was supported by STMP of Commission of Science and Technology of Shanghai (No.21XD1402500), and Shanghai Municipal STMP (2021SHZDZX0103).

References

[1] Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir Roshan Zamir. On evaluation of embodied navigation agents. *arXiv: Artificial Intelligence*, 2018. 6

[2] 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. *computer vision and pattern recognition*, 2017. 1, 2, 3, 6, 7

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[4] 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. 1, 2, 6, 7

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[6] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. 2

[7] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multitask mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 2, 3

[8] 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. *arXiv: Computer Vision and Pattern Recognition*, 2018. 1, 2, 7

[9] Steven W. Gangestad and Mark Snyder. Self-monitoring: Appraisal and reappraisal. *Psychological Bulletin*, 2000. 7

[10] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. *international conference on computer vision*, 2021. 6, 7

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146, 2020. 1, 2, 4, 7

[12] 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. 3

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[14] Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, and Stephen Gould. Language and visual entity relationship graph for agent navigation. *neural information processing systems*, 2020. 2

[15] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021. 1, 2, 3, 4, 6, 7

[16] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: Multimodal, multitask representations across languages. *empirical methods in natural language processing*, 2021. 3

[17] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022. 2

[18] Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah A. Smith, and Yejin Choi. Robust navigation with language pretraining and stochastic sampling. *empirical methods in natural language processing*, 2019. 1, 2, 7

[19] Xiangru Lin, Guanbin Li, and Yizhou Yu. Scene-intuitive agent for remote embodied visual grounding. *computer vision and pattern recognition*, 2021. 1, 2

[20] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *neural information processing systems*, 2020. 2, 3, 4, 7

[21] Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv: Artificial Intelligence*, 2019. 2

[22] 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. *european conference on computer vision*, 2022. 1, 2

[23] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *international conference on machine learning*, 2016. 6

[24] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The (un)surprising effectiveness of pre-trained vision models for control. *international conference on machine learning*, 2022. 3

[25] 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. *international conference on computer vision*, 2021. 1, 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *international conference on machine learning*, 2021. 2, 3, 6

[27] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetić, Mario Lučić, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *arXiv preprint arXiv:2206.06922*, 2022. 3

[28] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. *international conference on computer vision*, 2019. 2

[29] 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? *Learning*, 2021. 3

[30] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. *north american chapter of the association for computational linguistics*, 2019. 1, 2, 4, 7

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *neural information processing systems*, 2017. 1

[32] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. *computer vision and pattern recognition*, 2018. 2, 7

[33] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989. 6

[34] Fei Xia, Amir Roshan Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. *computer vision and pattern recognition*, 2018. 2

[35] Hong-Xing Yu, Leonidas J Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. *arXiv preprint arXiv:2107.07905*, 2021. 3

[36] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. *computer vision and pattern recognition*, 2018. 2

[37] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. *computer vision and pattern recognition*, 2020. 7

[38] Yifeng Zhuang, Qiang Sun, Yanwei Fu, Lifeng Chen, and Xiangyang Xue. Local slot attention for vision and language navigation. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 545–553, 2022. 3