AerialVLN: Vision-and-Language Navigation for UAVs

Shubo Liu†† Hongsheng Zhang†† Yuankai Qi Peng Wang† Yanning Zhang
Qi Wu∗
1Northwestern Polytechnical University
2University of Adelaide
{shubo.liu, hongsheng.zhang}@mail.nwpu.edu.cn, qykshr@gmail.com,
{peng.wang, ynzhang}@nwpu.edu.cn, qi.wu01@adelaide.edu.au

**Instruction:** Take off, fly through the tower of cable bridge and down to the end of the road. Turn left, fly over the five-floor building with a yellow shop sign and down to the intersection on the left. Head to the park and turn right, fly along the edge of the park. March forward, at the intersection turn right, and finally land in front of the building with a red billboard on its rooftop.

Figure 1: An intelligent agent should be able to follow given natural language instructions and navigate to the destination in an unseen environment with visual perceptions along the way. The green line shows the agent’s ground truth trajectory, and the chequered flag represents the end of it.

Abstract

Recently emerged Vision-and-Language Navigation (VLN) tasks have drawn significant attention in both computer vision and natural language processing communities. Existing VLN tasks are built for agents that navigate on the ground, either indoors or outdoors. However, many tasks require intelligent agents to carry out in the sky, such as UAV-based goods delivery, traffic/security patrol, and scenery tour, to name a few. Navigating in the sky is more complicated than on the ground because agents need to consider the flying height and more complex spatial relationship reasoning. To fill this gap and facilitate research in this field, we propose a new task named AerialVLN, which is UAV-based and towards outdoor environments. We develop a 3D simulator rendered by near-realistic pictures of 25 city-level scenarios. Our simulator supports continuous navigation, environment extension and configuration. We also proposed an extended baseline model based on the widely-used cross-modal-alignment (CMA) naviga-
1. Introduction

Recently, a bunch of vision-and-language navigation tasks, such as R2R [2], RxR [20], REVERIE [28], TouchDown [7], Alfred [33], iGibson [23, 32, 36], have drawn a large amount of attention from different research communities like computer vision, natural language processing and robotics. These tasks as well as their datasets have greatly boosted the research of assembling the capabilities of vision and language understanding, cross-modality matching, path planning and reasoning [6, 8, 15, 18, 27]. However, all these VLN tasks are designed for ground-based agents, which means agents can only navigate indoors or outdoors on the ground. This overlooks another important application scenario: activities in the sky, which are becoming increasingly popular with the development of unmanned aerial vehicles (UAVs), especially multirotor. We can now use UAVs to enjoy spectacular scenes without going out of houses and they can be potentially utilized for goods delivery, traffic surveillance, search/rescue and security patrol [10, 11, 12].

To release humans from manually operating UAVs and to fill the research gap in the field of navigation in the sky, we propose a city-level UAV-based vision-and-language navigation task, named AerialVLN, and a corresponding dataset. Navigating in the sky is significantly different from that on the ground in several aspects. First, AerialVLN has a larger action space. Compared to conventional ground VLN [2, 7, 19, 20, 28], AerialVLN requires intelligent agents to additionally take actions such as “rise up” and “pan down” into consideration. Moreover, multirotors can move left/right without turning its head. Second, the outdoor environments of AerialVLN are much bigger and more complex. AerialVLN covers a large variety of city-level scenes. An intelligent agent is required to distinguish referred buildings/objects by their spatial relationship from a bird-view as shown in Figure 1. Although the TouchDown [7] task is also devised for outdoor navigation, its environments are static, while ours are interactive and dynamic. For example, our agent can land on a building, and the weather and illumination conditions can dynamically change in the environment. Third, to mimic multirotor flying in real life, our AerialVLN has a much longer path than ground VLN’s. On average, our AerialVLN involves a path length of 661.8 units.4 There are about 9.7 referred objects in one instruction on average, which is more than 2.6 times as many as in the R2R dataset [2]. Fourth, intelligent agents must learn to avoid getting stuck on objects in 3D space. This is more challenging than avoiding obstacles when navigating on the ground as in VLN-CE [19] because agents have to estimate the 3D shapes of obstacles and the distance to obstacles. All these new characters render AerialVLN a different and highly challenging task.

AerialVLN is implemented using Unreal Engine 4 [14] and Microsoft AirSim plugins [31], which enables continuous navigation and near-realistic rendering. In total, we have collected 25 different city-level environments, covering a variety of scenes such as downtown cities, factories, parks, and villages, including more than 870 different kinds of objects. Our AerialVLN dataset consists of 8,446 flying paths obtained by experienced human UAV pilots who hold the AOPA (Aircraft Owners and Pilots Association) certificate. We pair each path with 3 instructions annotated by AMT workers in the standard dataset setting. Notably, we also align each sub-path to its sub-instruction, which enables fine-grained cross-modality matching learning. On average, up to 83 words are in each instruction, involving a large vocabulary of 4,470 words. Finally, we evaluate five baselines, including two golden standard VLN models in VLN, Seq2Seq model and cross-modal matching (CMA) model, and our proposed model to serve as starting baselines on AerialVLN.

2. Related Work

In this section, we review two types of closely related work: UAV navigation and Ground-based VLN.

UAV Navigation. Unmanned Aerial Vehicle (UAV) navigation has brought increasing attention in the last few decades. Early UAV autonomous navigation requires solving the challenges of perceiving, mapping, localisation, decision-making (path-planning), action-decomposing and controlling. Inertial UAV navigation and GPS-based methods are commonly used together since the former might cause significant errors due to accumulation, and the latter is usually unable to localise the vehicle in high precision [22, 26]. However, navigation in GPS-denied and unknown environments (such as cities with collapsed buildings or complex electromagnetic scenarios) becomes the bottleneck of intelligent UAVs. Vision-based navigation is then believed to be the solution to autonomous navigation [9, 24, 34].

The most similar works to ours are [3, 4, 13, 25] that language instructions are also provided. In [3, 4, 25], a

---

4 One unit equals one meter in our simulated city environment.
quadcopter agent is required to navigate by following natural language instructions in only one closed virtual field. The environment is of size 50×50 with 6×13 landmarks. It has 6,000 instructions with an average length of 57 words and a vocabulary size 2,292. Agents are only allowed for horizontal movements (Forward, Left/Right and Stop). By contrast, our AerialVLN is much larger, more complex, and closer to real-world scenarios: AerialVLN provides 25k crowd-sourced natural language instruction with an average length of 83 words and a vocabulary size 4,470. AerialVLN has 870 different kinds of objects and allows agents to move in 4-DOF as multirotor (Forward, Turn Left/Right, Ascent/Descend, Move Left/Right and Stop). Moreover, AerialVLN presents 25 different open city-level environments, which enables intelligent agents to be trained and tested more comprehensively. Compared to ANDH [13], which focuses on dialogue-based aerial VLN with bird-view image input, our AerialVLN task requires agents to navigate with a first-person view and our environments are interactive and dynamic, which requires agents to learn to avoid obstacles. Regarding path length and data amount, our AerialVLN is four times of ANDH. More comparisons can be found in Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R2R [2]</td>
<td>7,189</td>
<td>21,567</td>
<td>Indoor, discrete</td>
<td>Instruction</td>
<td>Graph-based</td>
<td>10.0</td>
<td>5</td>
<td>3.1k</td>
<td>29</td>
</tr>
<tr>
<td>RxR [21]</td>
<td>13,992</td>
<td>13,992</td>
<td>Indoor, discrete</td>
<td>Instruction</td>
<td>Graph-based</td>
<td>14.9</td>
<td>8</td>
<td>7.0k</td>
<td>129</td>
</tr>
<tr>
<td>CV2DN [35]</td>
<td>7,415</td>
<td>2,050</td>
<td>Indoor, discrete</td>
<td>Instruction</td>
<td>Dialog</td>
<td>25.0</td>
<td>7</td>
<td>4.4k</td>
<td>34</td>
</tr>
<tr>
<td>REVERIE [28]</td>
<td>7k</td>
<td>21,702</td>
<td>Indoor, discrete</td>
<td>Instruction</td>
<td>Graph-based</td>
<td>10.0</td>
<td>5</td>
<td>1.6k</td>
<td>18</td>
</tr>
<tr>
<td>SOON [37]</td>
<td>40k</td>
<td>3,848</td>
<td>Indoor, discrete</td>
<td>Instruction</td>
<td>Graph-based</td>
<td>16.8</td>
<td>9</td>
<td>1.6k</td>
<td>39</td>
</tr>
<tr>
<td>TouchDown [7]</td>
<td>9,326</td>
<td>9,326</td>
<td>Outdoor, discrete</td>
<td>Instruction</td>
<td>Graph-based</td>
<td>313.9</td>
<td>35</td>
<td>5.0k</td>
<td>90</td>
</tr>
<tr>
<td>VLN-CE [19]</td>
<td>4,475</td>
<td>13,425</td>
<td>Indoor, continuous</td>
<td>Instruction</td>
<td>2 DoF</td>
<td>11.1</td>
<td>56</td>
<td>4.3k</td>
<td>19</td>
</tr>
<tr>
<td>LANI [25]</td>
<td>6,000</td>
<td>6,000</td>
<td>Outdoor, continuous</td>
<td>Instruction</td>
<td>2 DoF</td>
<td>17.3</td>
<td>116</td>
<td>2.3k</td>
<td>57</td>
</tr>
<tr>
<td>ANDH [13]</td>
<td>6,269</td>
<td>6,269</td>
<td>Outdoor, continuous</td>
<td>Dialog</td>
<td>3 DoF</td>
<td>144.7</td>
<td>7</td>
<td>3.3k</td>
<td>89</td>
</tr>
<tr>
<td>CVDN [35]</td>
<td>7,415</td>
<td>2,050</td>
<td>Indoor, discrete</td>
<td>Instruction</td>
<td>Familiarity</td>
<td>266.8</td>
<td>204</td>
<td>4.3k</td>
<td>83</td>
</tr>
<tr>
<td>AerialVLN-S</td>
<td>8,446</td>
<td>25,338</td>
<td>Outdoor, continuous</td>
<td>Instruction</td>
<td>4 DoF</td>
<td>661.8</td>
<td>204</td>
<td>4.3k</td>
<td>83</td>
</tr>
<tr>
<td>AerialVLN</td>
<td>3,916</td>
<td>11,748</td>
<td>Outdoor, continuous</td>
<td>Instruction</td>
<td>4 DoF</td>
<td>321.3</td>
<td>115</td>
<td>2.8k</td>
<td>82</td>
</tr>
<tr>
<td>ANDH [13]</td>
<td>6,269</td>
<td>6,269</td>
<td>Outdoor, continuous</td>
<td>Instruction</td>
<td>Familiarity</td>
<td>144.7</td>
<td>7</td>
<td>3.3k</td>
<td>89</td>
</tr>
<tr>
<td>CVDN [35]</td>
<td>7,415</td>
<td>2,050</td>
<td>Indoor, discrete</td>
<td>Instruction</td>
<td>Familiarity</td>
<td>266.8</td>
<td>204</td>
<td>4.3k</td>
<td>83</td>
</tr>
<tr>
<td>AerialVLN-S</td>
<td>8,446</td>
<td>25,338</td>
<td>Outdoor, continuous</td>
<td>Instruction</td>
<td>Familiarity</td>
<td>661.8</td>
<td>204</td>
<td>4.3k</td>
<td>83</td>
</tr>
<tr>
<td>AerialVLN</td>
<td>3,916</td>
<td>11,748</td>
<td>Outdoor, continuous</td>
<td>Instruction</td>
<td>Familiarity</td>
<td>321.3</td>
<td>115</td>
<td>2.8k</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 1: Comparison of existing vision-and-language navigation tasks. AerialVLN presents a city-level open environment dataset for aerial vision-and-language instruction-based navigation. Note that the en-US subset of RxR is considered for a fair comparison. Path length unit: meter.

However, all the above-mentioned VLN tasks are designed for intelligent agents navigating on the ground, as shown in Table 1. This cannot reflect the challenges when navigating in the sky for intelligent agents like multirotor. To address this problem, in this work, we propose a new task AerialVLN, which is designed for UAV navigation in the sky.

3. The AerialVLN Task

As shown in Figure 1, the proposed AerialVLN task requires an intelligent agent (a multirotor in virtual environments) to fly to the destination by following a given natural language instruction and its first-person view visual perceptions provided by the simulator. Unlike previous VLN tasks [2, 20, 37], we do not provide pre-build navigation graphs in our task, so any point not occupied by objects (such as buildings and trees) is navigable. This is closer to the practical scenario.

Formally, at the beginning of each episode, the agent is placed in an initial pose $P = [x, y, z, p, r, y']$, where $(x, y, z)$ denotes the agent’s position and $(p, r, y')$ represents pitch, roll, yaw portion of the agent’s orientation. Then given a natural language instruction $X = \omega_1, \omega_2, ..., \omega_L >$, where $L$ is the length of instruction and $\omega_i$ is a single word token, the agent is required to predict a series of actions. The agent can take both the instruction and visual perceptions into consideration. Although our adopted simulator can provide panoramic observations, here we follow the most robotic navigation tasks [19] setting to limit our baseline agent to the access of its front view perceptions (both depth and RGB images) $V_t = \{v_t^R, v_t^P\}$. The agent needs to rotate to obtain other views. Navigation ends when the agent predicts a Stop action or reaches a pre-defined maximum action number. The navigation is recognised as a success if the agent stops at a location that is less than 20 units to the target location, as 20 metres is the common size of a
helipad in most countries. Considering the average size (radius) of our environment is around 3867.8 meters, this 20 metres landing area is rather small and challenging. The next section provides more details about the visual observation and action space.

4. Simulator

Our simulator is developed based on AirSim [31] and Unreal Engine 4 [14]. Below we detail its visual perceptions and action space.

Visual Observations. In the simulator, an embodied agent can move and observe in the continuous outdoor environment freely. At each step \( t \), the simulator outputs an RGB image \( v^R_t \) and a depth image \( v^D_t \) of its front view. Considering the outdoor environment setting, the depth sensor is allowed to perceive 100 meters ahead. In addition, semantic segments are also accessible for future usage. The simulator supports dynamic environments, such as blowing leaves, running cars, varying illumination (morning, noon, night) and different climate patterns (sun, rain, snow, fog). This can greatly narrow the gap when transferring trained agents to the real-world [1].

Action Space. Although the simulator supports flying towards any given direction and speed/distance, we consider the eight most common low-level actions in UAVs: \textit{Move Forward}, \textit{Turn Left}, \textit{Turn Right}, \textit{Ascend}, \textit{Descend}, \textit{Move Left}, \textit{Move Right} and \textit{Stop}. To balance the number of actions performed in a trajectory and the actual movement of the drone in an outdoor environment, the \textit{Move Forward} action continuously moves 5 units along its current direction. The \textit{Move Left} and \textit{Move Right} actions continuously move 5 units along the corresponding direction, respectively. The \textit{Turn Left} and \textit{Turn Right} actions turn 15 degrees horizontally. The \textit{Ascend} and \textit{Descend} actions continuously move 2 units vertically.

5. Dataset

In this section, we present the data collection policy and analysis of collected instructions for AerialVLN.

5.1. Data Collection

The data collection process contains two main steps: path generation and instruction collection. In contrast to R2R [2, 4, 5], of which the ground truth trajectories are automatically generated from navigation graph, we employed experienced multirotor manipulators (AOPA licensed) to complete the flying. This enables the trained agent to learn real human remote pilots’ behaviours. To ensure the high quality of paths with a reasonable length, we ask human manipulators to pass several randomly-selected landmarks from a pre-defined landmark set including buildings, fountains, squares etc. In the simulator, we also provide hints about directions and distance to the next landmark to manipulators, which can help them better accomplish the task (refer to the supplementary for the interface). The output of the path generation step includes the multirotor’s pose trace (a series of time-stamped 6-DoF multirotor poses). The raw flying paths may have redundant motions, as manipulators sometimes need to look around to identify their positions and decide where to go. We remove such redundant motion for smoother ground truth trajectories. Then, the continuous paths are discretised into meta actions, such as “turn left” and “move forward” to enable training.

For the second step, we use Amazon Mechanical Turk (AMT) to collect language instructions for these paths. Specifically, we show videos of drone flights and require the annotators to give natural language commands that can lead a pilot to complete the flying (refer to supplementary for the interface). To enrich the language diversity and reduce bias, each video is annotated three times by different annotators. To avoid ambiguity caused by similar landmarks, referring expression is required, such as “land on the rooftop of the building near fountain”. To validate data quality, all the collected instructions are manually checked by another group of workers. More details about the validation policy are presented in the supplementary material.

5.2. Data Analysis

We totally collected 25,338 instructions with a vocabulary of 4,470 words. On average, each instruction has 83 words. Figure 2 presents the relative word frequency in the form of word cloud, where the larger the font, the more frequently the word is used. Figure 2(a) shows that “building” and “road” are mostly used as reference objects for navigation. Figure 2(b) shows that “turn”, “go” and “fly” are the most common verbs.

In Table 1, we provide a comparison between our AerialVLN dataset and other popular VLN datasets. It shows that AerialVLN has the largest average path length, which is about five times ANDH and 40 times the ground-based VLN, RxR. At the same time, our dataset has the most average actions per path, which is about four times

Figure 2: Statistics of nouns and verbs.
Table 2: Linguistic phenomena in randomly sampled 25 instructions. AerialVLN task has a significant rate of reference, sequencing, spatial relationship and direction, which brings much more challenges to intelligent agents. p and μ represent the percentage of instructions that present the phenomena and the average number of the phenomena appears in each instruction. For a fair comparison, TouchDown Navigation subset is used.

VLN-CE and six times TouchDown. In terms of the number of instructions, our dataset is about twice VLN-CE and RxR, four times ANDH. All these characters render our dataset extremely challenging. In Figure 3, we present the distribution of instruction length and the number of actions. As shown in Figure 3(a), instruction length ranges from 50 to 130 words. Figure 3(b) shows that most paths have 50 ∼ 240 actions. These diversities make AerialVLN more challenging.

As in previous work [7, 20], we have also conducted statistics on linguistic phenomena of randomly sampled 25 instructions with comparison to other VLN datasets in Table 2. It shows that our AerialVLN task has a significant rate of reference, coreference, sequencing, spatial relationship and direction, which brings much more challenges.

**Dataset Split.** Following the common practice in the VLN community, we divide our dataset into the train, val_seen, val_unseen, and test splits. The word “seen” means the visual environments that have been seen in the train split. As shown in Table 3, we assign 17 scenes for training and val_seen split, where the train set contains 16,380 instructions from 5,460 paths. For val_seen, we assign 1,818 instructions from 606 paths. The val_unseen and the test split are both assigned 8 scenes, but the test split is about double the size of the val_unseen split, including 1610 paths with 4,830 instructions. Please note that the test split is built on unseen scenes as well and the goal locations for the test set will not be released. Instead, we provide an evaluation server where UAV trajectories can be uploaded for scoring.

Besides the standard dataset setting (with all 25 scenes), we also present a variant for small scenes, AerialVLN-S. It preserves the same split, but it has 17 scenes with a smaller scale and evenly-distributed path length, which results in a shorter path length (average path length reduces 51.5%) and shorter instruction length. In AerialVLN-S setting, the agent can sometimes even observe the goal at starting point. In this variant, there are 10,113 instructions for the train and 333 for validation seen splits, respectively. 531 instructions are assigned for validation unseen and 771 instructions for the test set, as shown in Table 3. We hope future researchers employ AerialVLN to tackle long path length 3D VLN tasks in the unseen environment and focus on the investigation of action learning in long time horizons and sparse reward; while utilising AerialVLN-S as the benchmark for general 3D aerial VLN tasks in first-person view.

Table 3: Dataset splits. AerialVLN is the full dataset, and AerialVLN-S is for small scenes.

6. Experiment and Results

In this section, we first present the evaluation metrics and training details of baseline models. Then we provide extensive evaluation and analysis.

6.1. Evaluation Metrics

We adopt four widely used metrics in VLN tasks [2, 17, 20, 28]: Success Rate (SR), where one navigation is
Figure 4: Main architecture of the Cross-Modal Attention model

considered successful if the agent stops within 20 meters of the destination; Oracle Success Rate (OSR), where one navigation is considered oracle success if the distance between the destination and any point on the trajectory is less than 20 meters; Navigation Error (NE), the distance between the stop location to the destination; Success rate weighted by Normalised Dynamic Time Warping (SDTW), which considers both the navigation success rate and the similarity between ground truth path and model predicted path [16].

6.2. Results

We evaluate five baseline models on our task. Four of these baselines have served as a golden standard for VLN tasks as in [1, 2, 28]. The other baseline is our extension to the best existing baseline. Below we first briefly introduce baselines and then present the results.

6.2.1 Baselines

Random. The agent randomly selects actions at each location and stops until the ‘stop’ action is selected or when reaching the max steps. This is widely used to reflect how big the solution space can be.

Action Sampling. Action Sampling agents explore the statistical characteristic of the dataset by sampling actions according to the action distribution of the training set. This can be used to measure the similarity of the action distribution on evaluation and training splits.

LingUNet. LingUNet [25] is a baseline model used by previously aerial VLN task LANI. Consider that LANI assume that agent can see the destination from the start point, LingUNet has an episode-wise paradigm. However, such assumption can’t stand in AerialVLN task, we thus adapt LingUNet model into a step-wise paradigm.

Sequence-to-Sequence. Seq2Seq [2] is a baseline model with a recurrent policy. It takes as input the concatenation of the RGB feature $\overline{v}_t^R = \text{meanpool}(\text{ResNet}_{RGB}(v_t^R))$, the Depth feature $\overline{v}_t^D = \text{ResNet}_{Depth}(v_t^D)$, and the instruction embedding $s = \text{LSTM}(\omega_1, \ldots, \omega_L)$. Then it projects them to a hidden representation $h_t^{(a)} = \text{GRU}([\overline{v}_t^R, \overline{v}_t^D, s, h_{t-1}^{(a)}])$, which is further used to predict a distribution over the action space. The one with the largest probability is selected as the next action $a_t = \arg\max_a \text{softmax}(W_a h_t^{(a)} + b_a)$.

Cross-Modal Attention. Cross-Modal Attention (CMA) is a classical baseline model for VLN tasks. As shown in Figure 4, the CMA baseline is based on a bi-directional LSTM and divides the whole process into two parts. One is tracking visual observations, and the other one is decision-making. The former is formulated as $h_t^{(attn)} = \text{GRU}([\overline{v}_t^R, \overline{v}_t^D, a_{t-1}, h_{t-1}^{(attn)}])$, where $a_{t-1} \in \mathbb{R}^{1 \times 32}$ and is a learned linear embedding of the previous action. The latter encodes the instruction embedding first and outputs the intermediate hidden state $S = \{s_1, \ldots, s_L\} = \text{BiLSTM}(\omega_1, \ldots, \omega_L)$. Then the attention mechanism is applied to instructions and images, which is $\hat{s}_t = \text{Attn}(S, h_t^{(attn)}), \hat{v}_t^R = \text{Attn}(v_t^R, \hat{s}_t), \hat{v}_t^D = \text{Attn}(v_t^D, \hat{s}_t)$, where Attn is a scaled dot-product attention. Finally, all these features and embeddings are concatenated to serve as the input of the recurrent network and predict an action for the agent to execute.

(a) Shortest path guidance (b) Look-ahead guidance

Figure 5: Illustration of Look-ahead Guidance. ‘A’ denotes starting location; ‘⋆’ denotes destination; ‘X’ denotes current location; Blue path denotes ground-truth; Yellow path denotes “generated ground-truth” when the agent deviates from the real ground-truth path.
Look-ahead Guidance (LAG). When training models in a student-forcing fashion, the ground-truth actions are usually determined according to the shortest path from the current location to the destination (see Figure 5(a)) in most existing methods. However, this is unreasonable because the instructions do not describe the shortest path from starting to the destination. To mitigate this issue, we inspired by [29] and propose a new strategy that generates ground-truth actions according to a “look-ahead” path. As shown in Figure 5(b), assuming the agent is at location X currently, the look-ahead path is determined by three steps: (1) find the shortest path to return to the ground-truth path (X→B in the example); (2) navigate along the ground-truth path 10 steps (look-ahead step = 10), assuming arrive at location C; (3) the look-ahead path is the shortest path from X to location C, and the ground-truth action for the next step is the first step on this path. We combine the aforementioned CMA model and our look-ahead guidance as our new baseline, denoted as LAG.

6.2.2 Results

Table 4 shows all the results. The results show that:

1. Random action hardly succeeds. The success rate of the Random model (Row 1 and S1) is 0%. Even if we sample actions according to the action distribution of the training split, the success rate still remains below 1% (Row 2 and S2). Moreover, the oracle success rate is always below 3% on unseen splits. This indicates an agent can hardly reach even passing by the destination if it cannot understand the instructions, visual perceptions, and their alignment.

2. LingUNet achieves limited success. Performance on Unseen cases only slightly better than Action Sampling (Row S2–S3). This may be attributed to the lack of recurrent structures in the decision component of the Baseline, resulting in the model’s inability to effectively model historical information.

3. The golden baselines Seq2Seq and CMA achieve success rate 1.0%~1.6% on unseen splits of the full dataset (Val_Unseen and Test_Unseen, Row 3~4) and 2.2%~3.9% on the AerialVLN-S dataset (Row S4~S5). At the same time, the oracle success rate also rises to 5% and 16%, respectively. This indicates that learning-based models have a larger chance to succeed than random models. However, the success rate is still rather low compared to human performance (SR: ~80%).

4. When applying the Dataset Aggregation (DA [30], an offline student-forcing strategy where executed actions are sampled from predictions instead of ground-truth actions) technique to mitigate the training-test disconnection problem (agents in test are not exposed to the consequences of their actions during training), the performance becomes better with about 6% improvement on seen split and about 1% improvement on unseen splits (Row S6~S7). This demonstrates exploring non-ground-truth actions helps learn more from training data and increases generalisation ability. On the other hand, comparing to the results of the same model (e.g., CMA-DA) on ground-based VLN tasks, such as continuous R2R, the performance on AerialVLN is rather low: SR 4.5% vs 27%. This indicates AerialVLN is much more challenging.

5. By incorporating our proposed look-ahead guidance (LAG) to the best baseline model CMA, it achieves further enhanced performance (Row S8) on unseen splits in terms of both SR and SDTW, which demonstrates the look-ahead guidance can help the agent to fly according to instructions.

We also present a qualitative result in Figure 6. It shows that when the agent can align visual and textual landmarks (as well as understand rotation commands), it
Possible reasons for failures We find that the length of path magnificently influences the success rate. Seq2Seq and CMA could follow instructions at an early stage but they cannot get back on track once deviate. Take AerialVLN for example, we further divide it into a long-path set (average path length 813.2m) and a short-path set (average path length 326.9m). Success rate on the former is only 1.8% while the latter can be up to 7.4%. Failure to stop correctly also leads to the low success rate. Supplementary material provides further failure analysis. As shown in Table 4, OSR of both Seq2Seq and CMA is significantly higher than SR, which suggests that the agent has passed the goal location and failed to stop around it.

6.3. Modality Ablation Study

To investigate the importance of different modalities in this task, we conduct an ablation study based on the CMA model via removing RGB, Depth, RGB+Depth (Vision), and Language from inputs, one by one. The results are presented in Table 5.

It shows that both the vision and language inputs play the most important role (Row 1 vs. Row 4, Row 1 vs. Row 5). This is reasonable because without either of them the task actually is non-sense. Additionally, the large performance drop indicates the dataset has little visual or textual bias. On the vision side, without depth information (Row 2) or RGB information (Row 3) leads to a success rate drop and without RGB drops more. This indicates both RGB and depth information matter to the final success and RGB information contributes more.

7. Conclusion

In this work, we introduce a new task and a large-scale dataset, AerialVLN, for the exploration of vision-and-language navigation in the sky. The linguistic analysis yields that AerialVLN dataset presents significant challenges to complex language understanding and its associated visual-textual alignment. We also evaluate
several widely adopted baselines, of which the performance drops significantly on our task and falls far behind human performance. This indicates that our task provides a broad study space for further research.

8. Acknowledgement

This work was supported by National Key R&D Program of China (No.2020AAA0106900), the National Natural Science Foundation of China (No.U19B2037), Shaanxi Provincial Key R&D Program (No.2021KWZ-03), and Natural Science Basic Research Program of Shaanxi (No.2021JCW-03).
References


[23] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hywon Gweon, Jiajun Wu, Li Fei-


