Elevating Aerial Vision-Language Navigation with enhanced feature extraction backbone using GELAN and PGI

Shaurya Kishore Panwar
Delhi Technological University, Delhi, India

Ajeet Kumar
Delhi Technological University, Delhi, India

Anil Singh Parihar
Delhi Technological University, Delhi, India

ABSTRACT: Aerial Vision and Dialog Navigation(AVDN) is an emerging field in the domain of multimodal learning and Vision-Language Understanding. In our work, we introduce a new feature extraction backbone using GELAN (Generalized Efficient Layer Aggregation Network) architecture and Programmable Gradient Information (PGI) [(Wang et al.(2024))], for capturing more comprehensive context within the images. The improved backbone is based on a dual-path strategy that optimizes gradient flow and feature reuse, significantly enhancing learning efficiency and inference speed without excessive computational costs. Our model is able to beat SOTA(State-Of-The-Art) performance on the widely used Aerial Vision Navigation from Dialog History (ANDH)(Fan et al.(2022)) Task, in multiple metrics.

Keywords: Vision-language Understanding and Navigation, Unmanned Aerial Vehicles, Multimodal learning, Computer Vision, Natural Language Processing, Large Vision-Language Models

1 INTRODUCTION

Recent advancements in learning to navigate with Visuo-Linguistic inputs in photo-realistic environments have sparked increasing research interest, offering fundamental insights into multimodal representations. Much of this research has focused on indoor navigation, with tasks like R2R(Anderson et al.(2018)) and RXR(Ku et al.(2020)) being prominent, some work is also being done in outdoor street navigation [Touchdown(Chen et al.(2019)), StreetNav(Jain et al.(2023))]. However, a significant breakthrough came with AVDN(Fan et al.(2022)) and AirVLN(Liu et al.(2023)), who proposed the task Aerial Vision Navigation from Dialog History (ANDH)(Fan et al.(2022)) at ICCV 2023, extending Vision-Language Navigation (VLN) to aerial scenarios. This task introduced a unique challenge: navigating through a complex, continuous environment with Three Degrees-of-Freedom (3DoF). Thus, owing to its complexity, the task has established itself as one of the most difficult challenges within the domain of VLN. This integration of natural language processing into aerial navigation is not just a technical advancement; it represents a shift towards a more human-centric approach, enabling intuitive and hands-free drone control. Such developments have broad applications, from food delivery to wilderness rescue, significantly lowering the barrier to drone operation.

Navigating with visual and linguistic cues requires a deep understanding of directives, visual feature recognition, and adaptive interactions. ANDH(Fan et al.(2022)) presents additional challenges for Cross-modal grounding compared to standard VLN due to longer path trajectories for the drone to follow and wider image FOV (field-of-view) combined with 3DoF freedom of movement. TG-GAT(Su et al.(2023)) first introduced fine-grained grounding abilities, although we found that the CSPDarknet53(Bochkovskiy et al.(2020)) based backbone was ineffective at

capturing long-range dependencies and global context within visual features along the path and was also susceptible to information loss through a deep visual backbone.

To address the above challenges we introduce in our work, Object-*Grounded-Feature Extraction Backbone*(3.3), a new feature extraction backbone which uses Programmable Gradient Information(PGI)(Wang et al.(2024)) to prevent information loss through deep network and GELAN (Generalized Efficient Layer Aggregation Network)(Wang et al.(2024)) architecture for capturing more comprehensive context within the images, giving our model a holistic and a fine-grained view of the scenes. In summary our contributions with our newly proposed architecture can be summarised as :

1. A new and improved feature extraction backbone for better feature representation in Aerial Vision tasks, incorporating the YOLOv9 architecture with GELAN and PGI.

2. Surpassing previous state-of-the-art models on the widely used Aerial Vision Navigation from Dialog History (ANDH)(Fan et al.(2022)) challenge, as evident by the higher scores in metrics such as SR (Success Rate), and GP (Goal Progress).

2 RELATED WORKS

Vision-and-language Navigation(VLN), vision-language navigation is an emergent sub-field in the domain of multi-modal learning and Vision-Language Understanding. It involves an agent, that must act in an unknown environment using visual observations and Linguistic instructions from a user. Aerial-Vision-Dialog-Navigation is an extension on previous work in VLN datasets such as R2R(Anderson et al.(2018)), Rx2(Ku et al.(2020)), which employed a navigation graph based approach. Next, major change came with (VLN-CE)(Krantz et al.(2020)), which used a reconstructed 3D continuous environment from Matterport3D(Chang et al.(2017)) dataset so the agent could use continuous actions for navigation, much more akin to real world.

Aerial and Outdoor VLN Beyond that, there have been some research work in a more complex outdoor setting, such as TouchDown dataset(Chen et al.(2019)), StreetNav dataset(Jain et al.(2023)) and the modified LANI(Blukis et al.(2019)) dataset. First proposed using VLN for drones and is similar to AVDN(Fan et al.(2022)), but AVDN(Fan et al.(2022)) features a more realistic simulation by using Xview(Lam et al.(2018)) dataset for photo-realistic aerial images and gave 3-DoF(Degree-of-Freedom) control for the agent as compared to much constrained earlier works. Hence the reason that we chose to work with AVDN(Fan et al.(2022)) simulator for our research. Previous work on this dataset such as TG-GAT(Su et al.(2023)) was the first to introduce target-grounded object detection prediction in the head to improve performance, in our work we build upon it and improve it with enhanced feature detection backbone.

3 METHODOLOGY

3.1 *Problem Formulation*

In the **ANDH**(Fan et al.(2022)) task, an agent is initialized in the simulator with a history of dialogs D_{past}. The goal is to complete a sub-trajectory during each dialog round. For each episode, we receive the current dialog D_{curr}, historical dialogues D_{past}, current RGB bird's-eye-view area V_t, view corners C_t, position coordinates P_t, and heading angle θ_t. The ground-truth sub-trajectory of length L consists of view areas $T = c_t^1, c_t^2, ..., c_t^L$, where c_t^g are the GPS coordinates of the view vertices. During evaluation, the center points of view areas form the navigation trajectory, with c_t^L as the target region. The predicted navigation is successful if the IoU between the predicted final view area c_f and the target c_t^L exceeds 0.4.

$$G = \begin{cases} c_{T_{i+1}}, & \text{if } T_{i+1} \neq L \\ Des, & \text{otherwise} \end{cases} \tag{1}$$

The goal area G depends on the current navigation time step, and $\langle c_t^1, \ldots, c_t^L \rangle$ is the ground truth view area sequence.

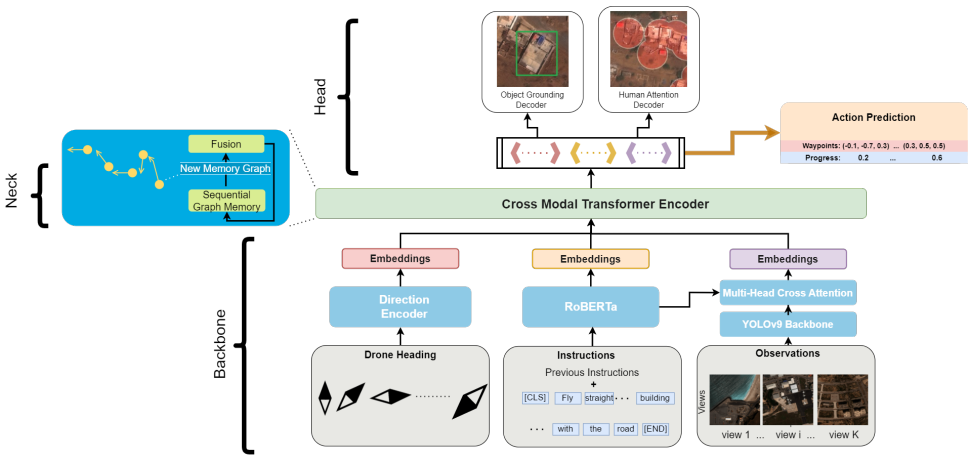

Figure 1. The given figure represents an overall overview of our pipeline, where Drone Heading, instructions and observations are first sent to their respective encoders to extract their feature embeddings which are then concatenated and passed through Cross-Modal Transformer Encoder, whose outputs are then passed to different heads(decoders) in the model which serve different purposes. For example: Object Grounding Head helps the agent learn fine-grained features in the observations, Human Attention Head helps the model build robust visuo-linguistic representations, Action Predictor Head is responsible for predicting drone motion (refer section 3.2).

Task	Routes	Instructions	Features	Language	Action Space	Actions
R2R (Anderson et al.(2018))	7,189	21,567	Indoor, discrete	Instruction	Graph-based	5
RxR (Ku et al.(2020))	13,992	13,992	Indoor, discrete	Instruction	Graph-based	8
CVDN (Thomason et al.(2020))	7,413	2,050	Indoor, discrete	Dialog	Graph-based	7
TouchDown (Chen et al.(2019))	9,326	9,326	Outdoor, discrete	Instruction	Graph-based	35
VLN-CE (Krantz et al.(2020))	4,475	13,425	Indoor, continuous	Instruction	2 DoF	56
LANI (Blukis et al.(2019))	6,000	6,000	Outdoor, continuous	Instruction	2 DoF	116
ANDH (Fan et al.(2022))	6,269	6,269	Outdoor, continuous	Dialog	3 DoF	7

Table 1. Comparison with ANDH Task with other tasks in Vision-Language Navigation. We can clearly see that ANDH presents much greater complexity than any other previous task.

3.2 *Overview*

As presented in figure 1, at each time step t our agent receives from the simulator, a combination of past dialogues D_{past} and current dialog D_{curr}, represented as $D = D_{curr} + D_{past}$, which are passed through an Large-Language-Model(LLM), RoBERTa(Liu et al.(2019)) to extract textual features ft_{tex}. The current bird's-eye-view V_t is passed through *Object-Grounded Feature Extraction Backbone* 3.3 to extract visual features ft_{vis}. Now, on the combined features ft_{tex} and ft_{vis} are passed to MHCA(Multi-Head Cross Attention mechanism) module from (Su et al.(2023)) to get final visual feature $\hat{ft_{vis}}$. In parallel, the drone headings are encoded by taking sine and cosines of the angles, which are then fed to a three-layer dense network to generate direction embeddings. Then the image and direction embeddings are stored in memory buffer for use in episodic graph memory encoder 2 similar to GAT(Su et al.(2023)). Finally, the concatenated embeddings are sent to Cross-Modal Transformer Encoder based on Episodic Transformer(Pashevich et al.(2021)), whose output is later sent to various decoder heads for action prediction, along with tasks like Object-grounding head(Su et al.(2023)) and Human Attention prediction(Fan et al.(2022)) that help the model learn robust visuo-linguistic correlation.

3.3 *Object-Grounded Feature Extraction Backbone*

To enhance the agent's capability of associating linguistic instructions with observable landmarks, TG-GAT(Su et al.(2023)) introduced Target-Grounding, a supplementary visual object-grounding task designed to forecast the bounding box of the referenced destination. The agent is trained to predict the bounding box $\hat{b} = [\hat{c}, \hat{x}, \hat{y}, \hat{w}, \hat{h}]$, where $\hat{c}$ represents the confidence score and $[\hat{x}, \hat{y}, \hat{w}, \hat{h}]$ denote the coordinates, within the input image that corresponds to the destination landmark described in the given language input. By explicitly learning to localize the referred

visual region, the agent strengthens its understanding of how linguistic references to landmarks are mapped onto the actual visual scene.

We introduced a major improvement in object-grounding by significantly improving feature-extraction backbone, we replace the traditional CSPDarknet53(Bochkovskiy et al.(2020)) backbone with updated YOLO-V9's(Wang et al.(2024)) backbone. The former followed a more traditional convolutional neural network (CNN) approach, with a mix of convolutional layers and Concentrated-Comprehensive Convolution (C3) modules, followed by spatial pyramid pooling(SPP). It struggled to capture long-range dependencies and global context effectively, our proposed architecture incorporates several advanced components that addressed these issues.

The upgraded Generalized Efficient Layer Aggregation Network (GELAN)(Wang et al.(2024)) architecture introduces a dual-path strategy that optimizes gradient flow and feature reuse, significantly enhancing learning efficiency and inference speed without excessive computational costs. It integrates Spatial Pyramid Pooling within the ELAN structure for multi-scale feature extraction, improving the ability to capture detailed and coarse image contexts across spatial hierarchies. Additionally, the new Programmable Gradient Information (PGI)(Wang et al.(2024)) addresses information loss during the feed-forward process, ensuring reliable gradients and enhancing model accuracy by maintaining crucial feature characteristics for more precise predictions. This comprehensive upgrade bolsters our object detection system's robustness and efficiency.

3.3.1 *Multi-Head Cross Attention(MHCA) mechanism*

The extracted contextual text embeddings ft_{tex} from RoBERTa(Liu et al.(2019)) along with flattened extracted visual features ft_{vis} from Object-Grounded Feature Extraction Backbone [Section 3.3] are passed to MHCA(Vaswani et al.(2017)) for feature aggregation as $\hat{F}_t$.

$$\hat{ft_{vis}} = \text{FFN}(I_{\text{cls}} + \text{MHCA}(ft_{vis})), \tag{2}$$

$$\text{MHCA}(ft_{vis}) = \text{Softmax}\left(\frac{I_{\text{cls}} W_q (ft_{vis} W_k)^T}{\sqrt{d}}\right) ft_{vis} W_v, \tag{3}$$

I_{cls} represents the contextual embedding of the [CLS] token, and W_q, W_k, and W_v are learnable matrices. Sine and cosine of heading angles are passed through a three-layer dense network to generate direction embeddings. The Multi-Head Cross-Attention (MHCA) output and direction embeddings are fed into a Cross-Modal Transformer Encoder, inspired by TG-GAT(Su et al.(2023)) and DUET(Chen et al.(2022)).

3.4 *Episodic Graph Encoding*

Our model incrementally builds a map using observations along the path, a concept introduced by DUET(Chen et al.(2022)) and expanded by TG-GAT(Su et al.(2023)), building on AVDN(Fan et al.(2022)). At each time step t, the agent receives concatenated embeddings C_t (image features $\hat{ft_{vis}}$, direction embeddings ft_{dir}, and text embeddings ft_{tex}), which are input into the Graph-Aware Transformer (GAT)(Su et al.(2023)). Since the agent also partially observes other nodes, we accumulate a combined representation of these navigable nodes based on the corresponding embedding in C_t. As depicted in Figure 2, GAT introduces an additional distance similarity G alongside the standard "Vision-Vision" query-key attention similarity values:

$$G = W_e E + b_e \quad ; \text{where } E = \|[x_i, y_i] - [x_j, y_j]\|_2 \tag{4}$$

W_e and b_e are learnable parameters, E_{ij} is the Euclidean distance between two locations, and ρ_i and ρ_j in Figure 2 share the same meaning. GAT replaces the conventional self-attention mechanism with a graph-attention approach.

4 RESULTS

Our proposed approach, which leverages better visual feature extraction backbone using GELAN(Wang et al.(2024)) and PGI(Wang et al.(2024)), which allowed our network to better capture both fine-grained and coarse-grained features within the input data, while making sure that important information is not lost in the deep feature extraction network.

Along with improving muli-modal understanding of our model, we were able to beat SOTA(State-Of-The-Art) performance on ANDH(Fan et al.(2022)) dataset, highlighting our model's superior

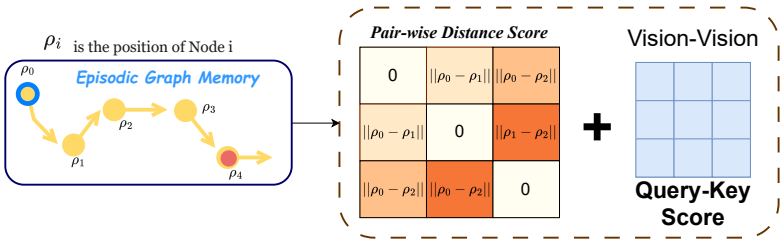

Figure 2. Representation of Graph Memory architecture, ρ_i represents historic previous locations and Vision-Vision Query-Key Scores are obtained from MHCA

visuo-linguistic understanding with an SR of 18.7 and GP of 59.8. We employed a variety of metrics to establish the efficacy of our model, such as SPL, SR and GP (please refer to 4.1). The utilization of multiple metrics ensures that our model not only achieves its goals but also does so reliably and efficiently.

	Seen Validation			Unseen Validation		
Model	SPL↑	SR↑	GP↑	SPL↑	SR↑	GP↑
HAA-LSTM(Fan et al.(2022))	11.6	13.0	50.3	18.3	20.0	54.4
HAA-Transformer(Fan et al.(2022))	14.7	17.3	56.3	16.5	20.4	55.2
TG-GAT(Su et al.(2023))	15.2	18.3	58.5	18.4	22.6	54.3
OURS	14.3	**18.7**	**59.8**	17.8	**23.3**	**57.9**

Table 2. The table presents results from the ANDH challenge, performance comparison of different models shows that the proposed model (OURS) achieved the highest Success Rate (SR) and Goal Progress (GP) scores in both Seen and Unseen Validation scenarios.

4.1 *Evaluation*

The predicted navigation is deemed a success when the Intersection over Union (IoU) between the forecasted terminal view region and the target area surpasses 0.4. Under these circumstances, multiple metrics are employed for assessment:

1. **Success Rate (SR)**: This is determined by calculating the ratio of successful predicted trajectories (those fulfilling the IoU criterion of exceeding 0.4) to the aggregate count of predicted trajectories.
2. **Success weighted by inverse Path Length (SPL)**: Originally put forth by (Anderson et al.(2018)), this metric gauges the Success Ratio weighted by the reciprocal of the path length, thereby imposing a penalty on more protracted trajectories.
3. **Goal Progress (GP)s**: As suggested by (Thomason et al.(2020)), this is derived by taking the Euclidean distance traversed by the agent along its trajectory and deducting the remaining distance from the midpoint of the anticipated target view area $\hat{\mathbf{c}}_N$ to the focal point of the goal region $\mathbf{G}$.

5 DISCUSSION & CONCLUSION

Since Aerial Vision and Language Navigation is a relative new and growing field we, see that there were few inconsistencies in AVDN(Fan et al.(2022)) dataset. Firstly, the authors' choice of target way-points outside the current Field-of-View (FOV) or along its edges introduces an unintended bias in the action distribution. This is particularly evident during the early stages of each navigation episode, where the majority of the drone's initial actions are concentrated along the periphery of its visual field. Additionally, there are instances of unintended noise in the dataset, such as mismatches between ground truth trajectories and ground truth target destinations.

Furthermore, since AVDN(Fan et al.(2022)) is based on XViews(Lam et al.(2018)), a 2D aerial image dataset, the actions of ascending or descending are not realistically portrayed. To address these issues, we plan to continue our research using AerialVLN(Liu et al.(2023)) , which utilizes a 3D simulation environment. Another goal is to deploy our model in adverse environments such as in low-light areas(Parihar et al.(2020)) and in presence of dynamic obstacles such as humans(Pandey and Parihar(2023)). This will provide a more realistic representation of real-life.

References

Anderson, Peter, Wu, Qi, Teney, Damien, Bruce, Jake, Johnson, Mark, Sünderhauf, Niko, Reid, Ian, Gould, Stephen, Van Den Hengel, Anton. 2018. "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3674–3683.

Blukis, Valts, Terme, Yannick, Niklasson, Eyvind, Knepper, Ross A, Artzi, Yoav. 2019. "Learning to map natural language instructions to physical quadcopter control using simulated flight." *arXiv preprint arXiv:1910.09664*.

Bochkovskiy, Alexey, Wang, Chien-Yao, Liao, Hong-Yuan Mark. 2020. "YOLOv4: Optimal Speed and Accuracy of Object Detection." *arXiv preprint arXiv:2004.10934*.

Chang, Angel, Dai, Angela, Funkhouser, Thomas, Halber, Maciej, Niessner, Matthias, Savva, Manolis, Song, Shuran, Zeng, Andy, Zhang, Yinda. 2017. "Matterport3D: Learning from RGB-D Data in Indoor Environments." In *International Conference on 3D Vision (3DV)*.

Chen, Howard, Suhr, Alane, Misra, Dipendra, Snavely, Noah, Artzi, Yoav. 2019. "Touchdown: Natural language navigation and spatial reasoning in visual street environments." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12538–12547.

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

Fan, Yue, Chen, Winson, Jiang, Tongzhou, Zhou, Chun, Zhang, Yi, Wang, Xin Eric. 2022. "Aerial Vision-and-Dialog Navigation." *arXiv preprint arXiv:2205.12219*.

Jain, Gaurav, Hindi, Basel, Zhang, Zihao, Srinivasula, Koushik, Xie, Mingyu, Ghasemi, Mahshid, Weiner, Daniel, Paris, Sophie Ana, Xu, Xin Yi Therese, Malcolm, Michael, et al. 2023. "StreetNav: Leveraging Street Cameras to Support Precise Outdoor Navigation for Blind Pedestrians." *arXiv preprint arXiv:2310.00491*.

Krantz, Jacob, Wijmans, Erik, Majumdar, Arjun, Batra, Dhruv, Lee, Stefan. 2020. "Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments." *arXiv preprint arXiv:2004.02857*.

Ku, Alexander, Anderson, Peter, Patel, Roma, Ie, Eugene, Baldridge, Jason. 2020. "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding." *arXiv preprint arXiv:2010.07954*.

Lam, Darius, Kuzma, Richard, McGee, Kevin, Dooley, Samuel, Laielli, Michael, Klaric, Matthew, Bulatov, Yaroslav, McCord, Brendan. 2018. "xView: Objects in Context in Overhead Imagery." *arXiv preprint arXiv:1802.07856*.

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, Stoyanov, Veselin. 2019. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692*.

Liu, Shubo, Zhang, Hongsheng, Qi, Yuankai, Wang, Peng, Zhang, Yanning, Wu, Qi. 2023. "AerialVLN: Vision-and-Language Navigation for UAVs." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15384–15394.

Pandey, Aman Kumar, Parihar, Anil Singh. 2023. "A Comparative Analysis of Deep Learning Based Human Action Recognition Algorithms." In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7. IEEE.

Parihar, Anil Singh, Singh, Kavinder, Rohilla, Hrithik, Asnani, Gul, Kour, Harpreet. 2020. "A comprehensive analysis of fusion-based image enhancement techniques." In *2020 4th international conference on intelligent computing and control systems (ICICCS)*, pp. 823–828. IEEE.

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

Su, Yifei, An, Dong, Xu, Yuan, Chen, Kehan, Huang, Yan. 2023. "Target-Grounded Graph-Aware Transformer for Aerial Vision-and-Dialog Navigation." *arXiv preprint arXiv:2308.11561*.

Thomason, Jesse, Murray, Michael, Cakmak, Maya, Zettlemoyer, Luke. 2020. "Vision-and-dialog navigation." In *Conference on Robot Learning*, 394–406. PMLR.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, Polosukhin, Illia. 2017. "Attention is all you need." *Advances in neural information processing systems*, 30.

Wang, Chien-Yao, Yeh, I-Hau, Liao, Hong-Yuan Mark. 2024. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information." *arXiv preprint arXiv:2402.13616*.