Open-Vocabulary Vision-Language Models for Autonomous UAV Obstacle Avoidance and Navigation

Faisal Ayub Khan, Soo Young Shin*

Department of IT Convergence Engineering, Kumoh national Institute of Technology, Gumi, South Korea
faisal@kumoh.ac.kr , *wdragon@kumoh.ac.kr

Abstract

Unmanned Aerial Vehicles (UAVs) operating in complex environments require not only accurate obstacle detection but also contextual understanding to ensure safe and efficient navigation. Traditional systems relying solely on geometric data often lack the semantic awareness necessary for dynamic and diverse scenarios. In this work, we propose a vision-language model (VLM)-driven perception framework that integrates open-vocabulary object recognition, monocular depth estimation, and multimodal semantic-spatial reasoning. Our system enables UAVs to identify, interpret, and assess various obstacle types in real time, distinguishing between static and dynamic entities such as buildings, pedestrians, and tree branches. Through training and evaluation on a custom UAV dataset, the model demonstrates robust generalization and effective decision-making in cluttered urban environments. The proposed approach establishes a bridge between low-level sensing and high-level autonomy, offering a scalable solution for intelligent UAV navigation.

1. Introduction

Unmanned Aerial Vehicles (UAVs) are increasingly deployed in complex environments where dynamic and semantically diverse obstacles pose significant challenges to autonomous navigation [1]. Traditional obstacle avoidance systems rely primarily on geometric data, lacking contextual understanding of the environment. This paper explores the integration of Vision-Language Models (VLMs) into UAV navigation systems to enhance real-time obstacle detection and avoidance through semantic mapping and contextual reasoning [2]. By leveraging VLMs, UAVs can not only detect but also interpret obstacle types—distinguishing, for example, between static structures like buildings and dynamic elements like pedestrians—and assess their impact on flight paths [3]. This semantic awareness enables more adaptable, efficient, and safe navigation strategies, particularly in cluttered urban or industrial environments. Potential applications of this approach include autonomous drones for infrastructure inspection, last-mile delivery, and exploration in environments where both obstacle density and semantic diversity are high. This work aims to demonstrate how VLM-driven perception can bridge the gap between low-level sensing and high-level decision-making in autonomous UAV systems.

2. System Model

Fig: [1] illustrates the architecture for proposed core concept. To enable robust and context-aware obstacle avoidance in UAVs, we leverage a vision-language model (VLM)-based semantic perception pipeline that integrates multimodal reasoning with spatial awareness. Given an input image I from the UAV's onboard RGB camera, a region proposal network (RPN) identifies candidate regions $R = \{r_1, r_2, ..., r_n\}$, which are subsequently processed through a pre-trained image encoder $f_I(.)$. In parallel, task-specific textual prompts—e.g., *"a photo of a [category] in the scene"* are embedded using a pre-trained text encoder $f_T(.)$. The model computes similarity scores between region features and text embeddings via dot product matching, enabling open-vocabulary object recognition, even for previously unseen categories. This semantic output $S = \{s_1, ..., s_n\}$ provides high-level contextual

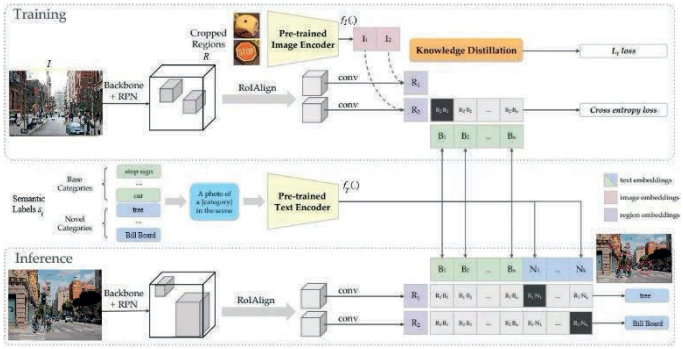

Figure 1: Overview of the proposed VLM-based perception architecture for UAV navigation

labels (e.g., "tree branch", "pedestrian") for each region. To localize obstacles in 3D space, the same image I is fed into a monocular depth estimation model $D(.)$, yielding dense depth maps $M \in \mathrm{R}^{H \times W}$. The resulting semantic-depth pairs (s_i, d_i) are passed to a reasoning module that performs risk assessment based on object type, motion status, and proximity. This joint semantic-spatial reasoning informs UAV navigation decisions in real time, enabling adaptive obstacle avoidance in complex environments.

2.1. VLM-Based Semantic Segmentation

To perform semantically informed scene understanding, we design a region-aware VLM segmentation module that integrates open-vocabulary recognition with multi-modal alignment [4]. Given an RGB image $I \in \mathrm{R}^{H \times W \times 3}$, a Region Proposal Network (RPN) extracts a set of candidate bounding boxes $R = \{r_i\}_{i=1}^{n}$ where each region r_i corresponds to a sub-image $I[r_i]$. Each region is projected into a high-dimensional embedding space via a vision encoder $f_I : \mathrm{R}^{h \times w \times 3} \rightarrow \mathrm{R}^d$, yielding region features $v_i = f_I(I[r_i]) \in \mathrm{R}^d$. Simultaneously, we encode a set of task-dependent natural language prompts $P = \{p_j\}_{j=1}^{m}$ such as "*a photo of a [category] in the scene*" using a language encoder $f_T : \text{Text} \rightarrow \mathrm{R}^d$. The semantic similarity matrix $S \in \mathrm{R}^{n \times m}$ is computed using normalized cosine similarity:

$$\mathbf{S}_{ij} = \frac{v_i^{\top} t_j}{\|v_i\| \cdot \|t_j\|}, \quad \forall i \in [1, n],\ j \in [1, m]. \tag{1}$$

To ensure robustness against semantically close classes, we apply temperature-scaled softmax normalization across categories:

$$P(s_i = j \mid r_i) = \frac{\exp(\mathbf{S}_{ij}/\tau)}{\sum_{k=1}^{m} \exp(\mathbf{S}_{ik}/\tau)}, \quad \tau > 0. \tag{2}$$

The most probable semantic class is assigned to each region:

$$s_i = \arg\max_j P(s_i = j \mid r_i), \tag{3}$$

enabling open-set recognition even in dynamically changing or previously unseen environments. This formulation allows UAV to semantically differentiate between objects like “pedestrian,” “tree branch,” or “construction crane,” based on textual prompts rather than fixed class labels.

2.2. Depth Estimation

Accurate spatial localization of semantic entities requires estimating depth from monocular imagery. We apply a dense regression-based depth estimator $D : \mathrm{R}^{H\times W\times 3} \rightarrow \mathrm{R}^{H\times W}$ to infer a continuous depth map $M = D(I)$, where each pixel $M(u,v) \in \mathrm{R}^{+}$ represents an estimated distance from the UAV to the corresponding point in the scene. For each region r_i, we extract depth features by computing the masked spatial average:

$$d_i = \frac{1}{|r_i|} \sum_{(u,v)\in r_i} M(u,v), \tag{4}$$

yielding a region-level depth estimate $d_i \in \mathrm{R}^{+}$. This allows pairing each semantic label s_i with a geometric localization d_i, forming semantic-depth tuples (s_i, d_i). For tasks requiring metric 3D localization (e.g., trajectory planning), the monocular depth map can be up-sampled and refined using stereo priors or SLAM integration modules

2.3. Multimodal Reasoning (Semantic + Depth Fusion)

To support real-time navigation, the UAV’s control module integrates semantic and spatial features via a reasoning function Φ that evaluates obstacle relevance and risk. Each scene element is represented by a triplet $z_i = (s_i, d_i, m_i)$, where s_i is the semantic label, d_i is the estimated depth, and $m_i \in \{0,1\}$ encodes motion status (static/dynamic), optionally derived from frame-differencing or optical flow. The risk score $r_i \in \mathrm{R}$ for each element is computed using a composite function

$$r_i = \Phi(s_i, d_i, m_i) = \alpha \cdot \psi_s(s_i) + \beta \cdot \psi_d(d_i) + \gamma \cdot \psi_m(m_i), \tag{5}$$

where ψ_s, ψ_d , and ψ_m are task-tuned basis functions representing semantic importance, proximity danger, and motion risk. respectively, and $\alpha, \beta, \gamma \in \mathrm{R}^{+}$. are tunable weights. The set of risk scores $\{r_i\}_{i=1}^{n}$ is then used to inform a navigation policy π: $\mathrm{R}^{n} \rightarrow A$, where A denotes the discrete action space:

$$a_t = \pi(\{r_i\}), \quad a_t \in \mathcal{A}. \tag{6}$$

This fusion enables the UAV to reason not just about where obstacles are, but ‘*what’* they are and *‘how’* they may affect its trajectory, leading to semantically-aware, dynamically-adaptive flight behaviors in complex environments.

3. Training and Evaluation

To validate the effectiveness of the proposed VLM-integrated perception pipeline, we trained the system using a custom dataset comprising 5,000 annotated RGB images captured from UAV flights in urban and semi-structured industrial environments. The vision-language module was initialized with a pre-trained CLIP backbone and fine-tuned using a contrastive loss over 30 epochs with a batch size of 32 and a learning rate of 1×10^{-5}, optimized via Adam. Depth estimation was trained om the UAV flight data, employing a ResNet-50 encoder- decoder architecture with a learning rate of 5×10^{-4} for 50 epochs.

During evaluation, full pipeline including region proposal, VLM-based semantic reasoning, and monocular depth estimation was tested on a held-out test set of 1,000 images. Quantitative results demonstrated accurate semantic segmentation (mean IoU of 0.68), reliable depth estimation (RMSE of 0.87 meters), and robust obstacle classification across a diverse set of scene categories. These findings confirm the model’s ability to generalize to unseen scenarios and support its potential for real-time UAV navigation in semantically complex environments.

4. Conclusion

This work presented a vision-language model-based perception framework for UAV navigation, enabling semantic understanding and context-aware obstacle avoidance. By integrating open-vocabulary recognition with monocular depth estimation and multimodal reasoning, the system achieves robust performance in complex, dynamic environments. Experimental results demonstrate the potential of VLM-driven perception to bridge the gap between low-level sensing and high-level decision-making in autonomous aerial systems

5. Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program(IITP-2025-RS-2024-00437190) supervised by the IITP (Institute for Information Communications Technology Planning Evaluation, 50%) This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (RS-2025-00553810, 50%)

References

[1] S. S. Y. Khan, F. A., “Real-time multimodal analysis for disaster man agement using uavs and vision-language model,” 2025 Korea Institute of Communications and Information Sciences Winter Conference Proceed ings, pp. 225– 226, 2025.

[2] Sautenkov, Y. Yaqoot, A. Lykov, M. A. Mustafa, G. Tadevosyan, A. Akhmetkazy, M. A. Cabrera, M. Martynov, S. Karaf, and D. Tset serukou, “Uav-vla: Vision-language-action system for large scale aerial mission generation,” in 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2025, pp. 1588–1592.

[3] K. Yuan, “Intelligent design and verification of aircraft autonomous obsta cle avoidance and collision avoidance system,” in 2024 IEEE 7th Interna tional Conference on Information Systems and Computer Aided Education (ICISCAE), 2024, pp. 603–607.

[4] Y. Guoand dB.Yang,“Asurveyofsemantic segmentation methods in traffic scenarios,” in 2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), 2022, pp. 452–457.