MorphoNavi: Aerial-Ground Robot Navigation with Object Oriented Mapping in Digital Twin

Sausar Karaf*, Mikhail Martynov*, Oleg Sautenkov, Zhanibek Darush, Dzmitry Tsetserukou

***Abstract*— This paper presents a novel object detecting and positioning approach for robotic systems utilizing a single monocular camera. The proposed system is capable of detecting a diverse range of objects and estimating their positions without requiring fine-tuning for specific environments. The system's performance was evaluated on the universal aerial-ground robot "MorphoGear" through a simulated search-and-rescue scenario, where the robot successfully located a robotic dog while an operator monitored the process. This work contributes to the development of intelligent, multimodal robotic systems capable of operating in unstructured environments.**

I. Introduction

Robotics has experienced rapid advancements in recent years, with Vision-Language Models (VLMs) emerging as a powerful tool for mission execution based on RGB images. Since VLMs require only an image and a text prompt as input, they eliminate the need for expensive and specialized sensors such as LiDARs and depth cameras. This simplicity and cost-effectiveness suggest that vision-language-based control will play a crucial role in the future of robotics, with cameras becoming the primary sensor for most robotic systems. This paper introduces a mapping method for a universal air-ground robot using a single camera. Our approach enables mission planning and easy VLM integration while reducing communication needs.

Unlike traditional approaches (point clouds, octomap [1], mesh) that preserve shape, our method retains semantic meaning. This enables high-level reasoning and rapid environment scanning. While less accurate, experiments (Section V) show precision is sufficient for global navigation, serving as a cost-effective alternative or LLM-augmentation for camera-based robots.

Our technique handles object recognition, segmentation, and depth estimation. By using common objects with standard dimensions as references, we can scale the map and estimate distances.

We prioritize semantic understanding over precise shape or color. Recognizing an object's purpose and context allows a robot to predict interactions, enabling robust navigation from minimal data, much like human reasoning.

II. Research Background

Monocular depth estimation is key for robotic 3D perception. *ZoeDepth* [2] advances this via hybrid relative and metric depth training. Similarly, *Depth-Anything* [3] offers a robust, zero-shot, scalable framework.

For detection, *YOLO* [4] and *Detectron2* [5] are fast but limited to predefined classes. In contrast, zero-shot detectors like *Grounding DINO 1.5 Pro* [6], *DINO X* [7], and *OWLV2* [8] recognize open-vocabulary objects, making them more suitable for our framework.

Image segmentation from models like the *Segment Anything Model (SAM)* [9] and *SAM v2* [10] provides object boundaries for depth scaling and planning. With *DINO 1.5* [11], these models enable training-free, semantic-aware perception, prioritizing high-level reasoning over geometric precision.

Unlike traditional mapping (e.g., SLAM point clouds), which loses semantics, modern zero-shot models retain spatial and semantic data efficiently. While VLMs like *Molmo* [12] and *ChatGPT* [13] integrate vision and language, they lack 3D reasoning. Vision-Language-Action models (*RT-1* [14], *PaLM-E* [15]) address this but need extensive task-specific data.

For path planning in aerial-ground robots (AGRs), *HE-Nav* [16] and *OMEGA* [17] offer solutions but require careful selection of safe take-off and landing points [18].

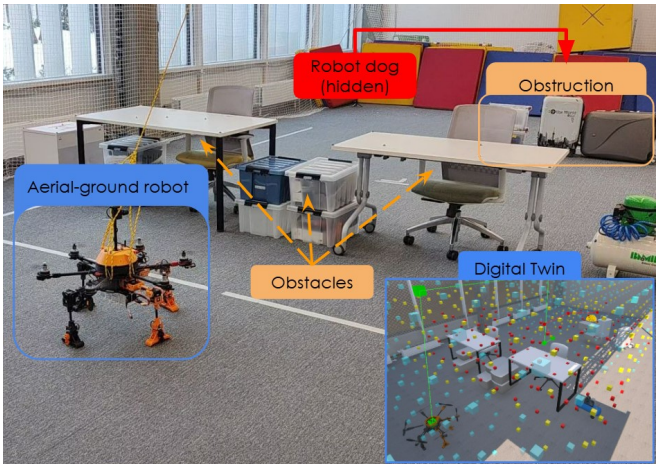

Fig. 1. Experimental setup. The mission is to overcome obstacles and search for the hidden robot.

III. System Overview

Our system consisted of three main parts, a robot, a laptop with a control interface, and an environment with a localization system. All calculations were performed either on the robot (for control) or on a personal computer (for mapping).

*Equal contribution

Authors are with Intelligent Space Robotics Laboratory, CDE, Skoltech, Bolshoy Boulevard, 30, bld. 1, Moscow 121205, Russia sausar.karaf, mikhail.martynov, oleg.sautenkov, zhanibek.darush, d.tsetserukou @skoltech.ru

Fig. 2. Aerial-Ground Vehicle MorphoGear.

A. MorphoGear: Aerial-Ground Robot

MorphoGear (Fig. 2) is an unmanned aerial-ground vehicle (AGV) with morphogenetic gear for terrestrial/aerial motion and future object manipulation. It overcomes non-traversable obstacles and can stop for data collection, enabling long-term operation impossible for drones.

Hardware includes an OrangePi 5b, OrangeCube flight controller, STM32-based limb controller, and ELP-USBFHD05H camera. Software uses ROS2 Iron with Python nodes and mavros. Limb motion generation continues previous work [19], moving scripts from Unity to internal asynchronous calculation. Ardupilot v4.4.1 runs on Ubuntu Server 22.04.

B. GUI: Ground Station

The ground station is a laptop (Intel i7-1165G7, 16GB RAM) running Unity and Python. We developed a digital twin for simulation and control, using ROS-TCP-Connector for ROS2 Iron communication. We built an experimental room in Unity with a robot digital twin featuring bidirectional communication (commands out, state in). The operator has free movement around the scene (Fig. 3).

1) Path planner: We developed an Aerial-Ground A* algorithm [20]. The interface visualizes nodes: red (occupied), yellow (dangerous), blue (free). Costs vary by layer: ground (none), first layer (high for takeoff/landing), upper layers (normal). A mission manager handles motion-type switching and node conflicts.

2) Object recognition script: A Python script processes single images for object recognition, segmentation, depth completion, and distance estimation, outputting object names and positions.

3) Object spawner: Detected objects are saved to JSON. An Object Spawner in Unity instantiates corresponding models from a library into the global frame, preserving semantic meaning and approximate dimensions while improving rendering performance.

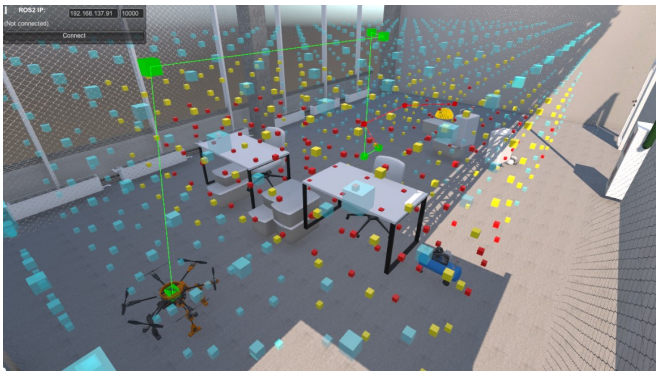

Fig. 3. Virtual simulation and visualization for MorphoGear.

C. Environment

Experiments were in a 6x10x4m room with a 5x8x3m planning grid, containing desks, chairs, suitcases, and other objects, using a VICON system for localization.

IV. Algorithm of Mapping

Effective navigation of a ground-aerial robotic system requires a map that accurately represents key environmental elements. The system must detect objects of interest and estimate their positions based on their known geometric dimensions. Once the 3D objects are positioned within the scene, the robot generates and follows a path to a user-defined destination.

The proposed system (Fig. 4) operates using a single monocular RGB image as input. During development, multiple object detection models, including OWLv2, OWL-ViT, and DINO-X, were evaluated. Among these, OWLv2 and Grounding DINO 1.5 Pro demonstrated the best performance in our testing environment (Section V) and were selected for implementation. Given the known object dimensions, camera intrinsics, and the bounding box obtained from the object detector, the object's distance is estimated using the following formula:

$$d = f\frac{h_m}{h_{px}}, \tag{1}$$

where f is the focal length, h_m is the object's real-world height (in meters), and h_{px} is its height in pixels.

To refine the distance estimates, we use Depth Anything v2 in combination with Segment Anything v2, leveraging a deep learning-based approach. A segmentation mask generated by Segment Anything is applied to the depth map, and the median depth value within the masked region is extracted. The final object distance is calculated as a weighted average of the two estimates: 80% from the geometry-based method and 20% from the depth-based method. When object dimensions are unknown, only the depth-based estimation is used.

The processed object positions are then packaged into a JSON file and transmitted to a Unity-based simulation environment. Within this virtual environment, 3D models of the detected objects are placed at their corresponding coordinates. A path planner then generates a trajectory through

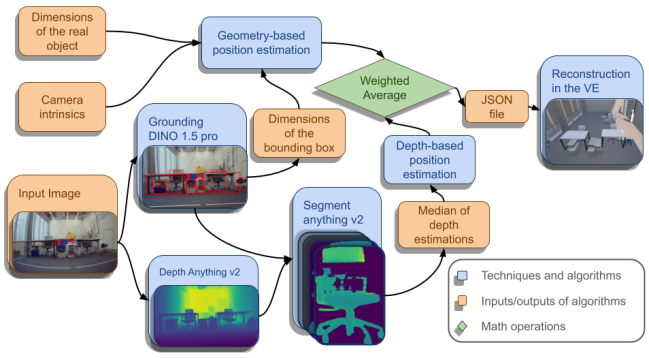

Fig. 4. The system architecture of the mapping pipeline.

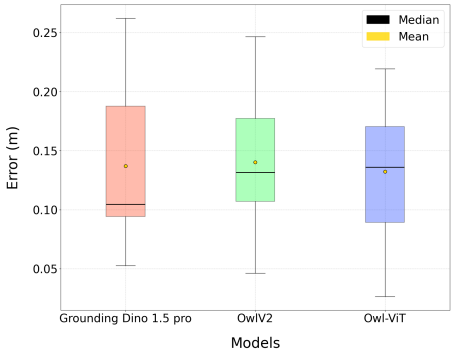

Fig. 5. Accuracy of position estimates.

the environment, which is subsequently sent to the robot. The onboard path tracking system of the aerial-ground robot follows this trajectory to reach the final destination.

V. Experiment

We evaluated our system in a simulated search and rescue scenario where the MorphoGear robot had to locate a Unitree Go1 robotic dog hidden behind obstacles. Performance was assessed via object detection ratio, position accuracy, and computation time.

A. Detection Model Choice

We evaluated state-of-the-art models (Table I). Grounding DINO 1.5 Pro achieved the best balance with a 97.4% detection ratio and moderate computation time, and was selected for experiments. Pose estimation accuracy (Fig. 5) confirmed this choice.

TABLE I
Comparison of Detection Models.

Model	Detection Ratio (%)	Calculation Time (sec)
Dino-X	90.4	7.26
Grounding Dino 1.6 pro	93.2	7.07
Grounding Dino 1.5 pro	**97.4**	**7.34**
Owl v2	69.5	13.21
Owl-ViT	37.5	4.77

B. Experimental Setup and Procedure

A test case evaluated system capabilities (Fig. 1). Obstacles were arranged linearly, dividing the area and hiding a robotic dog. The task required detecting obstacles, planning a path, navigating past them (validating locomotion transitions), and locating the target.

The procedure began with the robot capturing an image. The mapping pipeline processed this to compute object positions for the Unity GUI. An obstacle grid was constructed, and an A* algorithm planned a trajectory (Fig. 3). The pipeline (Fig. **??**) integrates depth and object size to improve pose estimation over noisy point clouds, enabling safer navigation.

C. Results and Limitations

The system detected 97.4% of target objects. The average object center position error was 13.6 cm compared to VICON ground truth. Scene reconstruction took 7.3 seconds on average. The robot successfully navigated obstacles to locate the target. Limitations include reduced accuracy for occluded objects, inaccuracies with irregular shapes/orientations, and a lack of real-time processing.

VI. Conclusion and Future Work

In this paper, we presented a mapping approach for a universal aerial-ground robotic system utilizing a single monocular camera. The proposed system demonstrated the ability to detect a diverse range of objects and estimate their positions without requiring fine-tuning for specific environments. Experimental validation was conducted through a simulated search-and-rescue scenario. The system achieved an object detection rate of 97.4%, with an average position estimation error of 13.6 cm and an average processing time of 7.34 sec per image.

While the system performed well in controlled laboratory conditions, several areas for improvement remain. Incorporating orientation estimation into the pipeline will lead to more accurate position calculations, particularly in cluttered environments. Additionally, occlusion remains a challenge, as partially visible or obstructed objects often result in incorrect position estimates. Future work will explore hierarchical and deep learning-based approaches to mitigate these issues.

Another crucial direction for future research is integrating the proposed mapping system with vision-language models (VLMs). By providing VLMs with structured scene information from our mapping pipeline in addition to the raw monocular image, we aim to enhance their spatial understanding and cognitive reasoning capabilities. This integration is expected to significantly improve the system's ability to interpret complex environments, leading to better decision-making in real-world applications.

Ultimately, this work contributes to the development of intelligent, multi-modal robotic systems capable of operating in unstructured environments. By addressing the identified limitations and expanding the system's capabilities, we move closer to achieving robust, autonomous aerial-ground navigation and perception for real-world deployment.

REFERENCES

[1] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: an efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[2] S. F. Bhat, I. Alhashim, and P. Wonka, "Zoedepth: Zero-shot transfer by combining relative depth and metric depth," in *2023 IEEE International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 12 345–12 350.

[3] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *2024 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 5678–5683.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788.

[5] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2: A versatile object detection and segmentation framework," 2024, gitHub repository.

[6] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, and J. Z. et al., "Grounding dino 1.5: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint,10.48550/arXiv.2405.10300*, 2024.

[7] T. Ren, Y. Chen, Q. Jiang, Z. Zeng, Y. Xiong, W. Liu, Z. Ma, J. Shen, Y. Gao, and X. J. et al., "Dino-x: A unified vision model for open-world object detection and understanding," *arXiv preprint, 10.48550/arXiv.2411.14347*, 2024.

[8] M. Minderer and A. G. et al., "Scaling open-vocabulary object detection," *arXiv preprint, 10.48550/arXiv.2306.09683*, 2023.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, and W.-Y. L. et al., "Segment anything," *arXiv preprint, 10.48550/arXiv.2304.02643*, 2023.

[10] A. Kirillov and E. M. et al., "Segment anything model v2: Scaling to new heights in zero-shot segmentation," *arXiv preprint, 10.48550/arXiv.2408.00714*, 2024.

[11] N. Carion and F. M. et al., "Grounding dino 1.5: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint,10.48550/arXiv.2405.10300*, 2024.

[12] M. Deitke, C. Schuhmann, Y. K. Shin, L. H. Li, R. Rombach, J. Hoffmann, I. Essa, and J. Liang, "Molmo: A family of open vision-language models," *arXiv preprint, 10.48550/arXiv.2409.17146*, 2024.

[13] OpenAI, "Chatgpt: Conversational ai powered by gpt architecture," OpenAI Technical Report, 10.48550/arXiv.2303.08774, Tech. Rep., 2022.

[14] A. B. et al., "Rt-1: Robotics transformer for real-world control at scale," in *arXiv preprint, arXiv:2212.06817*, 2022.

[15] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "Palm-e: An embodied multimodal language model," in *arXiv preprint, arXiv:2303.03378*, 2023.

[16] J. Wang, Z. Sun, X. Guan, T. Shen, D. Huang, Z. Zhang, T. Duan, F. Liu, and H. Cui, "He-nav: A high-performance and efficient navigation system for aerial-ground robots in cluttered environments," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 10 383–10 390, 2024.

[17] J. Wang, X. Guan, Z. Sun, T. Shen, D. Huang, F. Liu, and H. Cui, "Omega: Efficient occlusion-aware navigation for air-ground robots in dynamic environments via state space model," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1066–1073, 2025.

[18] X. Wang, K. Huang, X. Zhang, H. Sun, W. Liu, H. Liu, J. Li, and P. Lu, "Path planning for air-ground robot considering modal switching point optimization," in *2023 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2023, pp. 87–94.

[19] M. Martynov, Z. Darush, A. Fedoseev, and D. Tsetserukou, "Morphogear: An uav with multi-limb morphogenetic gear for rough-terrain locomotion," in *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, Seattle, WA, USA, 2023, pp. 11–16.

[20] M. A. Mustafa, Y. Yaqoot, M. Martynov, S. Karaf, and D. Tsetserukou, "Morphomove: Bi-modal path planner with mpc-based path follower for multi-limb morphogenetic uav," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Kuching, Malaysia, 2024, pp. 1820–1825.