Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00353-y

Original Article

Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study

Jegan Pranav [1], Emmanuel Joy [2, ✉], Suganthi Evangeline [3], Jenice Aroma [4]

1. *Artificial Intelligence and Machine Learning, Karunya Institute of Technology and Sciences, Coimbatore, IND*
2. *Department of M. Tech Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, IND*
3. *Department of Electronics and Communication Engineering, Sri Eshwar College of Engineering, Coimbatore, IND*
4. *Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, IND*

Received: June 12, 2026 | Review began: August 05, 2026 | Review ended: October 29, 2026 | Published: November 03, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

In the last five years, wars and worsening climate events have dramatically increased the necessity for Unmanned Aerial Vehicles (UAVs) using advanced Artificial Intelligence (AI) systems. The degree of usefulness that UAVs provide will depend on the onboard AI that can transform unprocessed imagery into actionable intelligence. This study reviews the developments in AI-based UAVs for identifying hazards and managing disasters from 2019 to 2025. The progression of deep learning architectures, such as YOLO to modern transformer architectures and Vision-Language Models (VLMs), has improved significantly recently. This study aims to propose a VLM-based pipeline for real-time hazard detection, reasoning, and situational reporting in a structured format. Experiments to evaluate detection performance in complex aerial scenarios using the VisDrone2019 dataset showed that the YOLOv11s lightweight architecture was superior to YOLOv11n (mAP at 50 = 0.51 vs mAP at 50 = 0.44). This study examines the role of multimodal sensing and reasoning to increase the quality of this system in complex environments. The challenges identified include real-time processing on edge devices, multimodal data fusion, and minimizing hallucinations in VLMs. The study ties together survey results with promising research directions for the future generation of UAV-based autonomous disaster response.

Categories: Computer Vision, Natural Language Processing (NLP)

Keywords: uavs, disaster response, yolov11, vision-language models, visdrone2019

Introduction

The increasing number and severity of natural and human-made disasters make it imperative that there be a rapid, safe, and accurate response to these events. Drones are now considered critical tools in today's disaster management efforts as they can be utilized for reconnaissance, Search & Rescue (SAR) operations, and assessing damaged infrastructure, in addition to identifying and evaluating the extent of damage from various types of disasters [1]. With an increasing number of disasters occurring globally at an accelerating rate, there is a great need for a prompt and effective disaster response that includes both a high degree of safety and access to large volumes of information generated by onboard intelligence. The large volumes of image-based data collected during disaster response operations require Artificial Intelligence (AI) techniques to convert raw imagery into useful actionable intelligence; therefore, it is also important to develop robust AI pipelines that will transform raw imagery into a structured environment that provides a relevant and

How to cite this article:
Pranav J, Joy E, Evangeline S, et al. (November 03, 2026) Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study. Cureus J Comput Sci 3 : es44389-026-00353-y. DOI https://doi.org/10.7759/s44389-026-00353-y

timely context to the operator. Beyond natural disasters, recent conflicts within the past five years (Russia/Ukraine, Gaza/Israel, and Iran/Iraq) have further demonstrated the utility of Unmanned Aerial Vehicles (UAVs) in reconnaissance and providing rapid situational awareness for humanitarian purposes. Similarly, escalating climate-related hazards, including flooding, wildfires, and earthquakes, have created a greater need for aerial monitoring utilizing drones. Collectively, these contexts highlight the potential of developing UAV systems that utilize AI for disaster and conflict response. Initial AI research related to UAVs has primarily focused on applying AI techniques to specific object detection tasks. For example, several researchers have applied convolutional neural networks (CNNs) (such as YOLO) to detect objects such as people, vehicles, and others. While these models are successful, they often cannot perform contextual reasoning, detect small objects, or deal with heavy occlusion [2]. Therefore, there is a clear need for AI models that possess a deeper level of contextual understanding and multiscale perception. Issues in previous models have been resolved using transformer-based models and hybrid architectures [3] that combine the efficiency of CNNs with awareness of global context. Fusing data from red, green and blue (RGB), thermal, and light detection and ranging (LiDAR) has increased robustness in adverse conditions. Recent advancements in transformers, hybrid CNN transformer architectures, and Vision-Language Models (VLMs) have aided UAVs in achieving semantic scene understanding, reporting, and reasoning in multifaceted disaster environments [4].

UAVs have changed significantly since the advent of VLMs and Large Language Models (LLMs). These models enable UAVs to detect objects, semantically understand environments, execute natural language commands, and generate human-readable reports [5]. This paper presents these developments and proposes an architecture of a VLM-based UAV pipeline made for hazard detection and reporting. In this study, a small experiment was conducted on the VisDrone2019 dataset, and the YOLOv11n and YOLOv11s variants were trained to demonstrate baseline detection in aerial scenes. This study combines various advancements to create a systematic description of the progression from CNN-based methods to LLM-guided autonomy. This research will also fill in this gap by linking the insights of the survey with the deployable UAV pipeline and real-world, practical systems of disaster response. This will make UAVs autonomous agents that are able to analyze hazards in context and provide actionable intelligent systems.

Background and related work

Previous Surveys and Reviews

Several surveys have observed AI applications in UAVs [6], and in a systematic literature review of 173 studies [7], researchers found that YOLO was the main framework for SAR and surveillance. Another study classified vision-based control and perception techniques, but disregarded multimodal sensing and transformer-based architectures, which are more applicable to UAV disaster responses. More recent studies have considered LLMs theoretically. Javaid et al. [6] studied the usage of GPT and BERT in UAV operations, and Lei et al. [8] broadly studied LLMs for disaster management. Despite these conceptual studies, they lack experimental validation and a real-world UAV focus. The UAV-focused evaluations that use VLMs remain rare, and few studies have delivered reproducible experimental validation using aerial datasets.

Research Gap and Motivation

Even with these valuable insights, existing solutions have three key limitations:

1. Temporal lag: Many surveys are outdated as the studies were conducted before the rise of VLMs, such as GPT-4o and LLaMA-3, and therefore do not cover the application of VLMs to UAV navigation.

2. Lack of synthesis: Most of the previous studies have either focused on detection using deep learning architectures or on theoretical frameworks for LLMs, but have not created a connection between existing foundational models and reasoning models.

3. Limited implementation works: Many evaluations, particularly those of existing works, do not evaluate implementations. Only a few studies have measured edge deployment constraints, real-time inference issues, and validated them using aerial datasets.

How to cite this article:
Pranav J, Joy E, Evangeline S, et al. (November 03, 2026) Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study. Cureus J Comput Sci 3 : es44389-026-00353-y. DOI https://doi.org/10.7759/s44389-026-00353-y

This survey bridges these gaps by combining the analysis of foundational object detection, transformer-based advances, and the pattern shift toward the VLMs and LLMs in real disaster scenarios.

Contributions of This Work

1. A unified survey from 2019 to 2025, which links CNN-based detectors, transformer evolutions, and VLM capabilities for UAV disaster responses by reasoning about and understanding the complex environments.

2. A conceptual pipeline designed for hazard detection and structured report generation, with clear use cases and evaluation methodologies.

3. A reproducibility box with experimental validation using the VisDrone2019 dataset with YOLO11n and YOLO11s models.

4. An outline of the proposed research agenda and a workflow of how to perform an evaluation that can be reproduced (datasets, baseline models, and metrics), so that empirical validation can occur.

5. An analysis of the real-world deployment challenges faced by UAV systems utilizing VLMs and a discussion of methods to ease those challenges.

Materials And Methods

Deep learning architectures for UAV disaster response

Real-time object detection is a core requirement for UAV applications [9], enabling fast identification of hazards and survivors. The YOLO family holds practical significance because of its efficiency and modest accuracy in aerial imagery [10,11]. Transformer and hybrid CNN transformer architectures improve contextual reasoning and small-target detection in complex scenes [12]. Edge-optimized transformer variants and quantization simplify onboard inference under computational constraints [13]. However, recent advances in VLMs and LLMs have extended UAV capabilities beyond detection.

CNNs and YOLO-based Approaches

The YOLO family has dominated UAV-based detection because of its real-time efficiency. However, baseline models degrade in performance at higher altitudes, speeds, or under occlusion. To address this, adaptations such as YOLOv5s with a feature fusion neck (PB-FPN) improved small-object detection, achieving a mean Average Precision (mAP) at 0.5 Intersection over Union (IoU), or mAP at 50, of 0.802 on SAR-specific datasets. In the context of post-disaster person detection, YOLOv3 with customized anchor boxes improves the accuracy by 46%. Other applications of YOLOv3 include suspicious activity detection, flood mapping, and body detection [14,15].

YOLO-style detectors have maintained a realistic trade-off between accuracy and inference speed for on-board applications, with variants that utilize multi-scale feature fusion and small-object detection being especially useful for aerial SAR applications. The test results depend on the dataset resolution and the UAV imagery's altitude or occlusion conditions. The following Table *1* gives a comparative analysis of YOLO variants on UAV SAR datasets.

Model	Metric Type	Reported Value	Key Contribution
YOLOv5s-PBfpn-Deconv [1]	mAP at 50	0.802	Small-object detection optimization
YOLOv3 (custom) [2]	AP30	0.60	Multi-stage training for SAR
YOLOv8 [10]	Accuracy	91%	High-altitude detection

PP-YOLO [14]	Accuracy	88.25%	Speed–accuracy balance
YOLOv11n	mAP at 50	0.44	Lightweight variant (VisDrone2019)
YOLOv11s	mAP at 50	0.51	Baseline validation (VisDrone2019)

TABLE 1: Performance comparison of YOLO variants on UAV datasets.

AP, Average Precision; mAP, Mean Average Precision; SAR, Search and Rescue; UAV, Unmanned Aerial Vehicle.

Transformer and Hybrid Models

Using self-attention, which encodes global context, small objects can be separated from background noise more effectively. For example, MLD-DETR, a DETR with multi-scale edge enhancement, achieved an AP50 score of 36.7% on the VisDrone2019 dataset, with a 20% reduction in parameters. Hybrid CNN transformers, such as DiRecNetV2, demonstrate the potential of both architectures, achieving a 0.964 F1-Score [176 frames per second (FPS)] for local and global information extraction. However, the multi-label disaster task achieves a much lower F1 Score (0.614), indicating that current models struggle to classify events in this task accurately. Edge-optimized transformers with post-training quantization can also classify disasters in real time on resource-constrained UAV platforms, enabling high-context perception.

Edge-adapted transformers and model compression techniques, such as pruning and quantization, are emerging trends that enable more deployable inferences on a wide range of devices with limited processing budgets. This has generated interest in making these techniques suitable for reasoning and reporting to assist autonomous UAVs in real-world disaster scenarios.

VLMs and LLMs for UAVs

VLMs include perception systems, as they convey semantics, adhere to instructions, and generalize to higher levels. Integrating detectors, contrastive encoders, and LLMs enables UAVs to interpret visual scenes and produce human-intelligible situational reports. Latency, hallucinations, and reliability are the key challenges that still exist.

VLM-Guided Navigation and Reasoning

In disaster scenarios with rapidly changing conditions, conventional waypoint-based navigation is inappropriate, and alternative methods, such as VLM-guided navigation [e.g., VLM-rapidly-exploring random tree (RRT)], can be useful. VLM-RRT [7] combines reasoning with GPT-4o and uses RRT to leverage VLM to guide the robot with directional clues with respect to the world. This fusion method halved the number of search iterations and demonstrated a 92% success rate in dynamically responding to moving targets in simulated wildfire scenarios.

Safety-aware methods like ASMA, which combines VLM environment encoders with MPC and scene-aware control barrier functions (CBFs), show a 64-67% improvement in success rate for real hardware quadrotor VLN tasks [16]. Safety-aware control methods that combine perception and CBFs could enable the integration of VLM outputs into verifiably safe and interpretable controllers. Realistic simulated environments with up to 12,000 trajectories were introduced by the OpenUAV benchmarks; large-scale pre-training, safety testing, and benchmarking of hierarchical LLM navigators [17] were enabled. In more realistic disaster environments, these benchmarks form a more critical basis. These are vital for evaluating the VLM-guided UAV autonomy.

Semantic Scene Understanding and Natural-Language Reporting

LLMs can improve UAV perception by generating descriptions of the identified objects. A pipeline using YOLOv7, CLIP, and GPT-3 produced real-time literary captions for UAV images and scored high on readability. These systems have limitations in their descriptive capacity, object-detection robustness, and ability to hallucinate details. Pipelines that integrate object detection, semantic encoders, and language models allow for automated report generation and mission summarization, bridging perception with human-interpretable analysis.

Surface-level evaluation metrics, such as BLEU and ROUGE, provide a limited measure of linguistic quality but fail to calculate operational utility in the context of UAV disasters. Therefore, standardized evaluation protocols, such as the LLVM-Drone [18], will be key to improving the reliability, interpretability, and synchronized assistance among UAV agents.

Multimodal sensing and benchmark datasets

Multimodal sensing, which combines RGB, thermal, LiDAR, and SAR, improves robustness under low-visibility conditions, such as smoke, dust, or nighttime conditions. Although UAVs can operate effectively in complex environments, data shortages for disaster phenomena and limited multimodal annotations remain major bottlenecks to robust model development.

Multimodal Fusion Approaches

RGB data alone is insufficient when encountering smoke, dust, or low-light conditions. Multi-sensor UAV systems detect hazards under all weather conditions by combining RGB, thermal, and LiDAR data. For instance, the levee inspection dataset used in the study [4] fuses thermal images to detect water seepage with LiDAR and RGB to compute water content using LiDAR echo intensity, achieving a 96.8% success rate during real flood disasters. DisasterM3 integrates SAR and optical data to train VLMs and assess global hazards [19]. These multimodal fusion architectures have a deeper understanding of occluded scenes with reduced visibility and can support disaster assessment tasks that could be challenging with a single-sensor architecture. Active efforts continue to develop lightweight sensor-fusion pipelines to be deployed on UAVs or drones.

The Critical Need for Specialized Datasets

While generic datasets like COCO and VisDrone are generally used for pre-training, the development of disaster response models continues to be limited by the lack of high-quality datasets for different DR scenarios, which remains a major bottleneck in the application of deep learning approaches in DR, especially in evaluating cross-domain generalization, scene reasoning, and multimodal grounding in UAV-based AI systems.

Datasets such as NOMAD are important, as occlusions are common in disaster areas, but recent results show that state-of-the-art models are not performing well on occluded images [20], so further research is needed in this area. Table *2* compares representative UAV and aerial datasets commonly used for disaster response, object detection, multimodal sensing, and emergency-scene understanding.

Applications in Disaster Management

The use of these systems for disaster response applications is growing [21]. UAVs with custom YOLO models have been used for SAR, fire detection, and flood monitoring. They have also been applied to damage assessment, victim localization, and real-time situation mapping using multimodal sensing inputs, such as RGB-thermal fusion and radar overlays.

Real-time inference on edge computing hardware, such as NVIDIA Jetson Nano, is a key enabler. An example is the YOLOv5s model, which, after optimization, can achieve approximately 2 FPS on the Jetson Nano, which is sufficient for low-speed aerial scanning.

Dataset Name	Modality	Size	Key Feature	Source
Real SAR Data	RGB	311 images	Real-world exercise data	[2]

DisasterEye	RGB (UAV)	2,751 images	Disaster classification	[13]
DisasterM3	SAR + Optical	26,988 images	VLM instruction tuning	[19]
NOMAD	RGB	42,825 frames	Systematic occlusion benchmark	[20]
VisDrone2019	RGB	10,209 images/frames	Aerial object detection benchmark	[22]

TABLE 2: Comparison of specialized UAV and aerial disaster response datasets.

RGB, Red, Green and Blue; SAR, Search and Rescue; UAV, Unmanned Aerial Vehicle; VLM, Vision-Language Model.

Future dataset designs should prioritize multimodal annotation, temporal consistency across modalities, and real-world disaster dynamics to enable reliable benchmarking of next-generation VLM and transformer-based UAV systems.

However, with recent resource-hungry architectures, such as transformers and VLMs, existing models exceed the capabilities of onboard devices. However, model optimization techniques, such as quantization and pruning or offloading to cloud/edge infrastructure, introduce compute latency and a network requirement. These factors make them prone to conditions such as communication-denied disaster scenarios. To overcome these limitations, hybrid computing strategies combine onboard inference for perception with offboard reasoning modules. Edge cloud collaboration frameworks, along with compression-aware model training, allow UAVs to maintain autonomy. They also enable UAVs to leverage external computing when available. For practical deployments, balancing autonomy, latency, and safety remains a key design challenge for UAV-based disaster systems, especially under unstable communication or power-constrained conditions.

Results And Discussion

Conceptual methodology - a VLM for hazard detection and reporting

Building on these advancements and addressing the limitations identified in the literature, we propose a VLM-based UAV system.

Advantages Over Prior Works

This methodology has several advantages over existing systems:

1. Actionable Intelligence: The model outputs prioritized, scenario-specific response recommendations (e.g., dispatch order, evacuation routing, resource allocation) rather than raw detections alone, so that human decision-makers can act on a ranked set of options instead of having to first interpret unstructured data. Descriptive systems mainly extract information from data, but they do not allow this.

2. Contextual Reasoning: By visualizing the scene, VLMs understand the situation by piecing together both spatial and semantic information. For example, by processing the full scene context, the VLM can identify not just a "person" but a "stranded individual" by reasoning that the person is on a rooftop surrounded by floodwaters. The VLMs can label them as "stranded" and not just a "person". It takes into account the whole scene, including the roof and the water, and can understand the person's situation.

3. Reduced Cognitive Load: The system can reduce the cognitive load on humans by not requiring them to analyze the video feeds and make decisions using the information gained from the video feeds.

Experimental validation and discussion

This section presents the experimental validation conducted using the YOLOv11n and YOLOv11s models on the VisDrone2019 dataset to establish a reproducible detection baseline for integration into the proposed VLM-based disaster-response pipeline. The goal was not to achieve state-of-the-art accuracy, but to provide a transparent, reproducible benchmark that informs subsequent multimodal and vision-language model experiments.

Experimental Setup

The YOLOv11n and YOLOv11s models were implemented using the Ultralytics YOLOv11 framework with a PyTorch backend and trained end-to-end from COCO-pretrained weights on a single-GPU, CUDA-accelerated workstation. Training followed the configuration summarized in Table *3* (optimizer, learning-rate schedule, batch size, and augmentation policy), with a fixed random seed used throughout to ensure deterministic, reproducible results. Indicative inference speed on Jetson-class edge hardware is reported qualitatively in the Results and Discussion section to assess real-time deployment feasibility.

Dataset and Preprocessing

We evaluated the models on the VisDrone2019 dataset, a public large-scale benchmark on aerial detection and tracking. We performed the experiments using both on the official training and validation splits. During these experiments, we resized the images to 640 by 640 pixels (for the baseline model) and 960 by 960 pixels (for high-resolution validation). For data augmentation, horizontal flipping, mosaic augmentation, scaling jitter, and random cropping were applied at different altitudes and illuminations. The model configuration and training parameters are shown in Table *3*.

Configuration	Details
Model Variants	YOLOv11n, YOLOv11s
Input Resolution	640 × 640, 960 × 960
Optimizer	AdamW, SGD
Learning Schedule	Cosine/Step LR; 100 epochs baseline
Batch Size	8
Loss Functions	Objectness, classification, CIoU box loss
Augmentations	Horizontal flip, mosaic, color jitter, scale jitter
Determinism	Fixed random seed for reproducibility
Metrics	mAP at 50 (primary), precision, recall, F1-score

TABLE 3: The Model configurations, training parameters, and evaluation metrics used in training YOLO models on the VisDrone2019 dataset.

The results demonstrate that YOLOv11s has a reasonable trade-off between performance and inference speed, making it suitable for use in real-time UAV applications, as shown in Figure *2* and Figure *3*. The YOLOv11n model is also well-suited for lightweight deployment on Jetson-class edge hardware, while maintaining an acceptable inference speed of 2 FPS for

low-speed aerial scans. As a component of the VLM-based disaster response pipeline, the detector provides bounding box output, class probability output, and uncertainty estimate output, providing a structured visual input for semantic reasoning and reporting capabilities. Future studies will evaluate the effectiveness of VLMs using fine-tuning, BLEU/ROUGE assessments of linguistic quality, and expert review of the interpretability and actionable nature of the models' outputs in disaster situations. The training metrics and accuracy for the model are depicted in Figure *2*.

Results and Discussion

In this section, we detail the main experiments conducted to evaluate the visual perception layer's performance as the data acquisition module of the potential VLM-based disaster response pipeline. YOLOv11 and its two variations, YOLOv11s and YOLOv11n, were tested on the VisDrone2019 dataset, with a focus on detecting small-scale objects such as pedestrians and vehicles in the highly cluttered aerial scene. We further evaluate our qualitative detection robustness in real-world settings, provide quantitative training metrics, and discuss trade-offs between detection accuracy and edge deployment.

Qualitative results of the model test in a drone video are shown in Figure *1*, where the trained YOLOv11s model is tested in a real-world environment through a series of top-down aerial frames captured to detect small objects. These results show a high degree of robustness of the model in detecting human beings and vehicles, despite the complexity of the variety of challenging real-world variables, such as ego-centric motion blur, rapid perspective shifts, and environmental vibrations, which the model successfully navigates to maintain consistent detection accuracy. In addition to detecting object instances, the system can aggregate this information to provide an estimate of crowd density and vehicle traffic flow in real time. This successful transition from static training sets to real-world scenario performance confirms that the visual perception layer can serve as a dependable backbone for the entire architecture. The successful operational performance of this system in the wild suggests that the visual perception layer is a reliable data input layer for the semantic reasoning layers of the proposed VLM pipeline.

The model has a high detection accuracy for high-density objects in difficult aerial scene images, including persons and vehicles at low and high altitudes, respectively. Despite the small object scale in the VisDrone dataset (compared with the general scene), the model provides tight bounding box location and low uncertainty for hard detections. This object-level perception will then serve as the "visual vocabulary" used by a VLM Core to perform higher-level semantic reasoning and hazard analyzes.

Training and validation performance for the different YOLOv11 variants are shown as metric curve plots in Figure *2* (YOLOv11s) and Figure *3* (YOLOv11n). Both the plots display steady performance through the full 100+ epoch training cycle with consistently decreasing objectness and classification loss values. Overall, the YOLOv11s variant performed best overall, achieving a mAP at 50 of 0.51 compared to the YOLOv11n variant that was able to achieve a mAP at 50 value of 0.44. This enhanced performance can be attributed to the greater number of parameters associated with the YOLOv11s variant, allowing for more robust feature extraction capabilities within complex aerial environments. Precision and recall plots demonstrate that both variants rapidly increase in performance during the first 20 epochs, after which they stabilize; supporting the use of the AdamW/SGD optimizer and cosine learning rate schedule to produce effective performance in the aerial environment. These results further provide a necessary operational tradeoff to support the use of YOLOv11 variants in real-time applications: While the YOLOv11s variant offers improved detection depth, the YOLOv11n variant has the ability to offer improved lightweight deployment characteristics with an inference time of 2 FPS on Jetson-class edge computing devices. Therefore, the YOLOv11n variant is likely to serve as a prime candidate for UAV-based disaster response applications where resources may be limited.

How to cite this article:
Pranav J, Joy E, Evangeline S, et al. (November 03, 2026) Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study. Cureus J Comput Sci 3 : es44389-026-00353-y. DOI https://doi.org/10.7759/s44389-026-00353-y



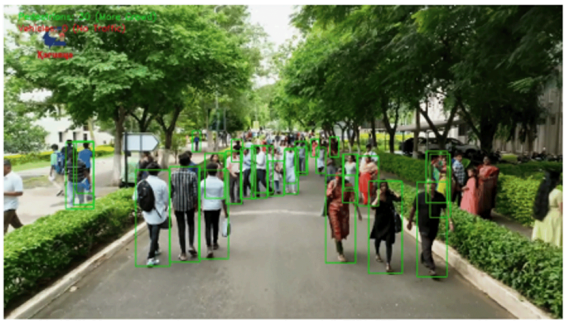
(a) Small-scale pedestrian localization in aerial view (Vehicle: None, Crowd: Low)



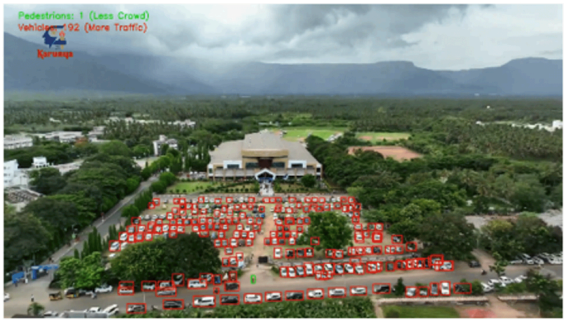
(b) Occluded vehicle detection under dense tree cover (Vehicle: Low, Crowd: None)



(c) High-density aerial crowd detection (Vehicle: None, Crowd: High)



(d) Cross-angle pedestrian detection along a narrow pathway (Vehicle: None, Crowd: High)



(e) Multi-vehicle detection in a congested parking area (Vehicle: High, Crowd: Low)



(f) Multi-class detection of pedestrians and vehicles (Vehicle: High, Crowd: High)

FIGURE 1: Sample detection from YOLOv11s illustrating the model's performance under varying conditions of vehicle density, crowd intensity, occlusion, scale, and viewpoint changes.

How to cite this article:
Pranav J, Joy E, Evangeline S, et al. (November 03, 2026) Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study. Cureus J Comput Sci 3 : es44389-026-00353-y. DOI https://doi.org/10.7759/s44389-026-00353-y

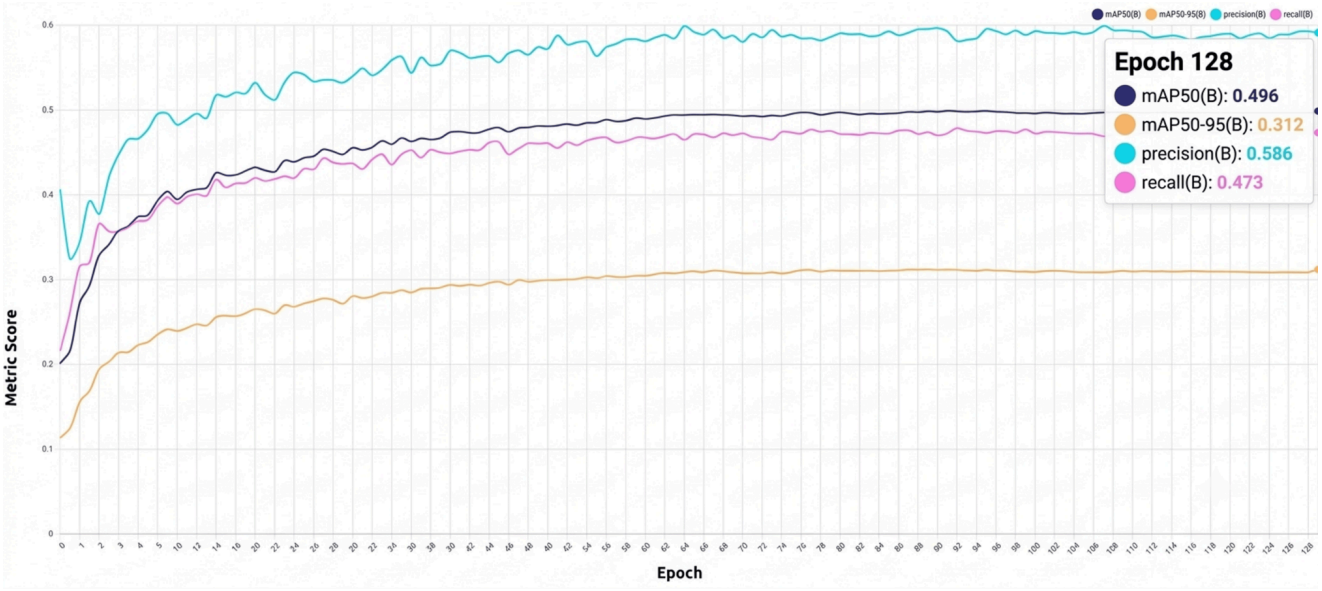


FIGURE 2: Metric curves for the YOLOv11s model.

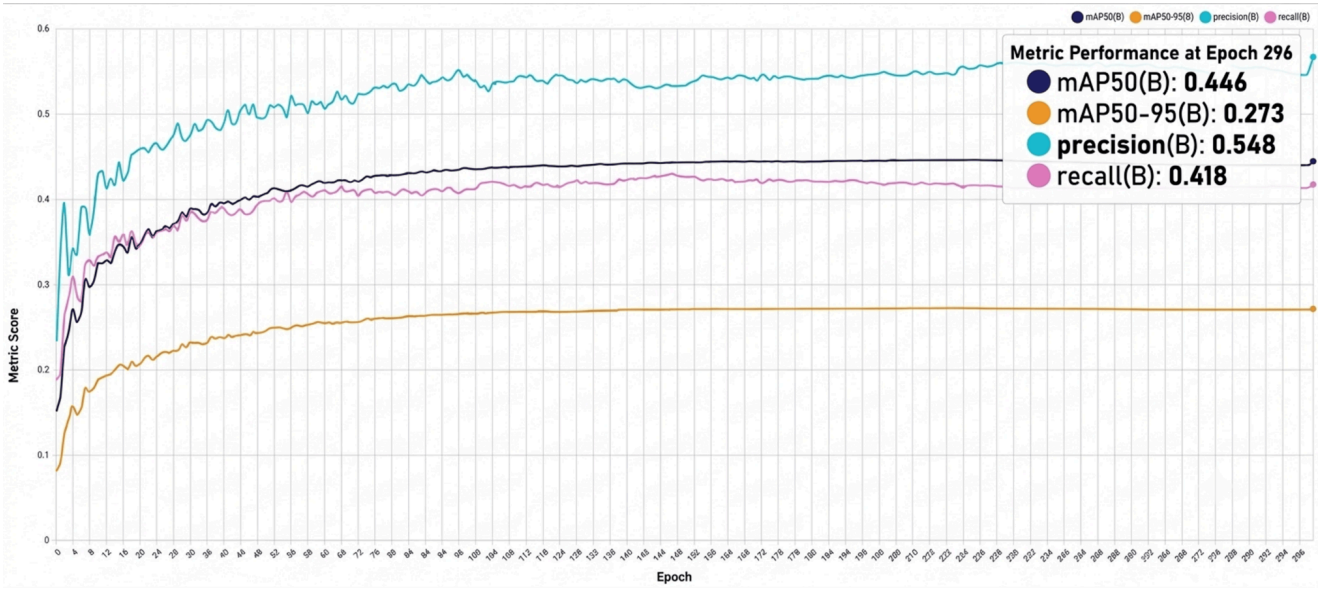


FIGURE 3: Metric curves for the YOLOv11n model.

Conclusions

Recent developments in AI-enabled UAVs were studied as they relate to their application in disaster response, ranging from CNN-based object detection techniques to transformer and VLMs for situational hazard assessment and reporting. A conceptual pipeline using a VLM was proposed for situational hazard detection and reporting, with reproducible baseline results from YOLOv11 tested against the VisDrone2019 dataset. Challenges remain in deploying these systems at the edge, integrating multiple modalities of sensor data, ensuring the trustworthiness of these models, and ensuring the operational reliability of these systems when operating under constrained conditions.

How to cite this article:
Pranav J, Joy E, Evangeline S, et al. (November 03, 2026) Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study. Cureus J Comput Sci 3 : es44389-026-00353-y. DOI https://doi.org/10.7759/s44389-026-00353-y

Future research directions include the optimization of VLMs for use in real-time onboard applications, extension of reasoning capabilities beyond RGB sensor data to include multimodal input data (RGB, thermal, LiDAR), and reduction of hallucinations to ensure that decision support can be both understood and reliable. Ultimately, the goal is to transform UAVs from simple passive sensors into fully autonomous cooperative agents that can provide safe, contextually aware, and human-centered disaster response.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Jegan Pranav, Emmanuel Joy

Acquisition, analysis, or interpretation of data: Jegan Pranav, Suganthi Evangeline, Jenice Aroma

Drafting of the manuscript: Jegan Pranav

Critical review of the manuscript for important intellectual content: Emmanuel Joy, Suganthi Evangeline, Jenice Aroma

Supervision: Emmanuel Joy

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request. The authors acknowledge that the original code and project supporting this study are hosted in the DroneX Surveillance System GitHub repository, publicly accessible at https://github.com/dremmjoy/DroneX-Surveillance-System. The VisDrone2019 dataset used for training and evaluation is likewise publicly available [22].

Acknowledgements

This paper is accepted and presented in International Conference on Sustainable Smart Computing and Communications (ICSSCC 2026) Organized In-person and Online (Hybrid Mode) by School of Computer Science and Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, India.

References

1. Ciccone F, Ceruti A: Real-time search and rescue with drones: a deep learning approach for small-object detection based on YOLO. Drones. 2025, 9:514. 10.3390/drones9080514
2. Bahmanyar R, Merkle N: Saving lives from above: person detection in disaster response using deep neural networks. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2023, X-1/W1-2023:343-350. 10.5194/isprs-annals-x-1-w1-2023-343-2023
3. Shianios D, Kolios PS, Kyrkou C: DiRecNetV2: a transformer-enhanced network for aerial disaster recognition. SN Computer Science. 2024, 5:770. 10.1007/s42979-024-03066-y
4. Su S, Yan L, Xie H, Chen C, Zhang X, Gao L, Zhang R: Multi-level hazard detection using a UAV-mounted multi-sensor for levee inspection. Drones. 2024, 8:90. 10.3390/drones8030090
5. de Curtò J, de Zarzà I, Calafate CT: Semantic scene understanding with large language models on unmanned aerial vehicles. Drones. 2023, 7:114. 10.3390/drones7020114
6. Javaid S, Fahim H, He B, Saeed N: Large language models for UAVs: current state and pathways to the future. IEEE Open Journal of Vehicular Technology. 2024, 5:1166-1192. 10.1109/ojvt.2024.3446799
7. Katkuri AVR, Madan H, Khatri N, Abdul-Qawy ASH, Patnaik KS: Autonomous UAV navigation using deep learning-based computer vision frameworks: A systematic literature review. Array. 2024, 23:100361. 10.1016/j.array.2024.100361
8. Lei Z, Dong Y, Li W, Ding R, Wang QR, Li J: Harnessing large language models for disaster management: A survey. Findings of the Association for Computational Linguistics: ACL 2025. 2025, 14528-14551. 10.18653/v1/2025.findings-acl.750
9. Ye J, Papaioannou S, Kolios P: VLM-RRT: vision language model guided RRT search for autonomous UAV navigation. 2025 International Conference on Unmanned Aircraft Systems (ICUAS). 2025, 633-640. 10.1109/ICUAS65942.2025.11007837
10. Maaroof MKA, Bouhlel MS: Real-time object detection using YOLO-8 model: a drone-based approach. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. 2025, 16:190-204. 10.58346/jowua.2025.i1.011
11. Thota JR, Padala A: Human remains detection in natural disasters using YOLO: a deep learning approach. Engineering, Technology & Applied Science Research. 2024, 14:17678-17682. 10.48084/etasr.8483
12. Zhu X, Zhang Z: Efficient vision transformers with edge enhancement for robust small target detection in drone-based remote sensing. Frontiers in Remote Sensing. 2025, 6:1599099. 10.3389/frsen.2025.1599099
13. Jankovic B, Jangirova S, Ullah W, Khan LU, Guizani M: UAV-assisted real-time disaster detection using optimized transformer model. 2025 IEEE Symposium on Computers and Communications (ISCC). 2025, 1-7. 10.1109/ISCC65549.2025.11326447
14. Polimera AJ, Parimi AM: Detection of suspicious activities at remote locations by using UAVs and computer vision. 2023 First International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV). 2023, 1-6. 10.1109/ICPEEV58650.2023.10391907
15. Hussain A, Latif G, Alghazo J, Kim E: Flood detection using deep learning methods from visual images. AIP Conference Proceedings. 2024, 3034:030004. 10.1063/5.0194669
16. Sanyal S, Roy K: ASMA: an adaptive safety margin algorithm for vision-language drone navigation via scene-aware control barrier functions. IEEE Robotics and Automation Letters. 2025, 10:9232-9239. 10.1109/lra.2025.3592138
17. Wang X, Yang D, Wang Z, et al.: Towards realistic UAV vision-language navigation: platform, benchmark and methodology. Proceedings of the 13th International Conference on Learning Representations (ICLR 2025). 2025, 75433-75451.
18. Xiao J, Zhang R, Zhang Y, Feroskhan M: Vision-based learning for drones: a survey. IEEE Transactions on Neural Networks and Learning Systems. 2025, 36:15601-15621. 10.1109/tnnls.2025.3564184
19. Wang J, Xuan W, Qi H, et al.: DisasterM3: a remote sensing vision-language dataset for disaster damage assessment and response. Advances in Neural Information Processing Systems 38 (NeurIPS 2025 Datasets and Benchmarks Track). 2025, 1-15.
20. Bernal AMR, Scheirer W, Cleland-Huang J: NOMAD: a natural, occluded, multi-scale aerial dataset for emergency response scenarios. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024, 8569-8580. 10.1109/wacv57701.2024.00839
21. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q: Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022, 44:7380-7399. 10.1109/tpami.2021.3119563

How to cite this article:
Pranav J, Joy E, Evangeline S, et al. (November 03, 2026) Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study. Cureus J Comput Sci 3 : es44389-026-00353-y. DOI https://doi.org/10.7759/s44389-026-00353-y

22. Mishra B, Garg D, Narang P, Mishra V: Drone-surveillance for search and rescue in natural disaster. Computer Communications. 2020, 156:1-10. 10.1016/j.comcom.2020.03.012

How to cite this article:
Pranav J, Joy E, Evangeline S, et al. (November 03, 2026) Integrating Deep Learning and Vision-Language Models for Unmanned Aerial Vehicle (UAV)-Based Disaster Response: A Survey and Experimental Study. Cureus J Comput Sci 3 : es44389-026-00353-y. DOI https://doi.org/10.7759/s44389-026-00353-y