A Multimodal Artificial Intelligence-Based UAV Patrol System for Intelligent Field Monitoring and Decision Support

Kuo-Lung Chen*

Department of Computer Science and Information Engineering, National Taitung University, Taiwan; e-mail: longerger14@gmail.com

*Corresponding Author: longerger14@gmail.com

DOI: https://doi.org/10.30211/JIC.202604.006

Submitted: May. 12, 2026 Accepted: Jul. 04, 2026

ABSTRACT

Unmanned aerial vehicles (UAVs) have been widely applied in field inspection, environmental monitoring, infrastructure management, and industrial patrol tasks due to their high mobility, flexible deployment, and capability to collect large-scale aerial imagery. However, conventional UAV patrol systems mainly rely on manual image interpretation or single-modal computer vision models, which limits their ability to understand complex field conditions, integrate heterogeneous information, and support intelligent decision-making. To address these challenges, this study proposes a multimodal artificial intelligence-based UAV patrol system that integrates aerial image acquisition, visual understanding, semantic analysis, and intelligent decision support. The proposed system combines UAV-based remote sensing, visual language models, object detection, image retrieval, and knowledge-driven reasoning to transform patrol images into structured and interpretable information. Through multimodal analysis, the system can identify field objects, detect abnormal conditions, understand spatial relationships, and provide natural-language responses for field management queries. The proposed framework is designed to support applications such as industrial site inspection, infrastructure monitoring, green energy field assessment, and smart campus patrol. Experimental evaluation demonstrates that the system can effectively improve patrol efficiency, reduce manual inspection workload, and enhance the interpretability of UAV-based monitoring results. The results indicate that multimodal artificial intelligence provides a promising approach for developing next-generation UAV patrol systems with visual perception, semantic understanding, and decision-support capabilities.

Keywords: Multimodal artificial intelligence, UAV patrol system, Visual language model, Remote sensing, Intelligent inspection, Decision support, Semantic reasoning

1. Introduction

With the rapid development of unmanned aerial vehicles (UAVs), remote sensing, edge computing, and artificial intelligence, UAV-based patrol systems have become an important technology for intelligent field monitoring and inspection. UAVs provide flexible deployment, low operational cost, high mobility, and the ability to collect high-resolution aerial images from large-scale or difficult-to-access environments. Therefore, UAVs have been widely applied in industrial site inspection, infrastructure monitoring, precision agriculture, disaster assessment, traffic surveillance, environmental protection, and smart city management.

In recent years, deep learning has significantly improved the visual perception capability of UAV-based systems. Wu et al. [1] reviewed deep learning-based UAV object detection and tracking methods and indicated that UAV imagery has become an important research topic across computer vision and remote sensing due to its flexible data acquisition capability and diverse application scenarios. However, UAV image analysis still faces several challenges, including small object detection, complex backgrounds, scale variations, occlusion, illumination changes, and different viewing angles. These challenges limit the reliability of conventional UAV patrol systems when they are deployed in complex real-world environments.

Traditional UAV patrol systems usually rely on manual image interpretation or task-specific computer vision models. Although object detection, image segmentation, and scene classification methods can identify predefined objects or abnormal patterns, they often lack semantic understanding and explainability. For example, a conventional model may detect vehicles, roads, buildings, or solar panels, but it is difficult for the model to explain spatial relationships, infer field conditions, or answer management-oriented questions. As a result, field managers still need to manually review a large number of UAV images and integrate their domain knowledge to make decisions.

To overcome the limitations of single-modal image analysis, visual question answering and vision-language learning have recently attracted increasing attention in the remote sensing community. Lobry et al. [2] proposed RSVQA for remote sensing visual question answering and demonstrated that natural-language questions can be used to extract meaningful information from remote sensing images. This research direction shows that remote sensing image analysis is gradually moving from simple visual recognition toward semantic understanding and human-computer interaction. Furthermore, recent studies have introduced transformer-based and question-driven methods to improve remote sensing visual question answering, enabling more complex reasoning over aerial images [3], [4].

More recently, large vision-language models and multimodal foundation models have provided new opportunities for intelligent UAV patrol systems. Weng et al. [5] reviewed vision-language modeling in remote sensing and pointed out that vision-language models can bridge the information gap between remote sensing images and natural language. These models support image captioning, visual question answering, image-text retrieval, and conversational interaction. In addition, EarthGPT was proposed as a universal multimodal large language model for multisensor image comprehension in the remote sensing domain [6]. These studies indicate that multimodal artificial intelligence can enhance the ability of remote sensing systems to understand complex visual scenes, integrate heterogeneous information, and provide interpretable responses.

Although existing IEEE studies have made significant progress in UAV perception, remote sensing visual question answering, and multimodal image understanding, several research gaps remain. First, most UAV patrol systems focus on visual detection rather than integrated semantic reasoning and decision support. Second, many remote sensing vision-language models are designed for satellite or general remote sensing images, while practical UAV patrol scenarios require closer integration with field management rules, inspection knowledge, and spatial relationships. Third, existing systems often lack a complete workflow that connects UAV image acquisition, multimodal understanding, semantic retrieval, and knowledge-driven decision support.

To address these issues, this study proposes a multimodal artificial intelligence-based UAV patrol system for intelligent field monitoring and decision support. The proposed system integrates UAV-based aerial image acquisition, visual perception, vision-language understanding, semantic image retrieval, and knowledge-driven reasoning. Through multimodal analysis, the system can identify important field objects, describe patrol scenes, detect abnormal conditions, retrieve relevant images using natural-language queries, and generate interpretable decision-support suggestions.

2. Literature Review

2.1 UAV-Based Patrol and Remote Sensing Image Analysis

UAV-based patrol has become an important research direction in remote sensing, smart inspection, and intelligent field management. Compared with traditional manual inspection, UAVs can rapidly acquire high-resolution aerial images from large-scale, dangerous, or difficult-to-access environments. Therefore, UAVs have been widely used in environmental monitoring, precision agriculture, traffic surveillance, infrastructure inspection, disaster assessment, photovoltaic field monitoring, smart campus management, and industrial site inspection.

With the rapid development of deep learning, UAV image analysis has gradually shifted from manual interpretation to automated visual perception. Wu et al. [7] provided a comprehensive survey of deep learning-based UAV object detection and tracking methods and indicated that UAV imagery has become an important research topic across computer vision and remote sensing. Their study summarized the challenges of UAV-based object detection and tracking, including small object size, complex backgrounds, motion blur, occlusion, and scale variation. These issues are highly related to practical UAV patrol scenarios because patrol images are often collected from different altitudes, viewpoints, and lighting conditions.

In addition to UAV-specific studies, deep learning has also been widely adopted in general remote sensing image analysis. Zhu et al. [8] reviewed deep learning methods for remote sensing and demonstrated that convolutional neural networks and representation learning have significantly improved the performance of image classification, object detection, semantic segmentation, and change detection. Cheng and Han [9] further reviewed object detection in optical remote sensing images and pointed out that remote sensing objects usually appear with large scale differences, arbitrary orientations, and complex spatial distributions. These characteristics are also common in UAV patrol images, especially when the system is deployed for field inspection and infrastructure monitoring.

Recent UAV object detection surveys have further emphasized that UAV images introduce additional difficulties compared with ordinary remote sensing images. Tang et al. [10] summarized recent advances in UAV object detection and noted that UAV-based aerial images are widely used in monitoring, geological exploration, precision agriculture, and disaster warning. However, due to low-altitude acquisition and complex field environments, UAV images often contain dense objects, background interference, and significant viewpoint changes. Therefore, robust UAV patrol systems require not only object detection capability but also semantic interpretation and contextual reasoning.

For practical patrol applications, object detection alone is usually insufficient. A UAV patrol system may detect buildings, vehicles, roads, solar panels, pipelines, fences, or vegetation, but field managers still need to understand the meaning of these detected objects. For example, a detected vehicle may be normal if it appears in a parking area, but it may become abnormal if it blocks an emergency route. Similarly, vegetation may be acceptable in a green area but may indicate a safety risk if it is close to power facilities. Therefore, UAV patrol systems need to move from object-level perception toward scene-level understanding and decision-oriented interpretation.

Although existing UAV image analysis methods have achieved promising results, most of them are still designed for predefined visual tasks, such as object detection, tracking, classification, or segmentation. These methods usually provide limited explanations and cannot flexibly answer user questions. Moreover, they are not designed to integrate patrol images with field knowledge, inspection rules, and management requirements. This limitation motivates the development of multimodal AI-based UAV patrol systems that combine visual perception, language understanding, semantic retrieval, and knowledge-driven decision support.

2.2 Multimodal Vision-Language Understanding for Remote Sensing

To overcome the limitations of single-modal visual recognition, multimodal vision-language learning has recently become an important research direction in remote sensing. Unlike traditional computer vision models that mainly output labels, bounding boxes, or segmentation masks, vision-language models can connect visual content with natural-language descriptions. This capability enables image captioning, visual question answering, image-text retrieval, visual grounding, and conversational image interpretation.

Remote sensing visual question answering is one of the representative tasks in this research direction. Lobry et al. [11] proposed RSVQA, which introduced visual question answering into remote sensing image analysis. Instead of requiring users to select predefined visual tasks, RSVQA allows users to ask natural-language questions about remote sensing images and obtain answers related to land cover, object counting, spatial relationships, and image content. This approach provides a more flexible way to access remote sensing information and reduces the dependence on domain experts for image interpretation.

Following RSVQA, several studies attempted to improve the reasoning ability of remote sensing visual question answering. Zhang et al. [12] proposed a multistep question-driven visual question answering method for remote sensing images, which enhances the ability of the model to decompose and answer complex questions. Zhang et al. [13] further introduced a spatial hierarchical reasoning network for remote sensing visual question answering, emphasizing the importance of spatial relationships in answering scene-related questions. These studies are highly relevant to UAV patrol because many practical inspection questions are spatially dependent, such as whether an obstacle is near a road, whether a facility is close to a restricted area, or whether an abnormal object appears around key equipment.

Vision-language models have also been adopted for image-text retrieval and semantic representation learning in remote sensing. Li et al. [14] developed deep cross-modal embedding networks for zero-shot remote sensing image scene classification, showing that cross-modal representation can help connect visual scenes with semantic descriptions. This type of image-text representation is important for UAV patrol systems because patrol missions may generate a large number of images, and users need to search relevant images using natural-language queries rather than only file names, timestamps, or coordinates.

More recently, large vision-language models and multimodal large language models have provided new opportunities for remote sensing image understanding. Weng et al. [15] reviewed vision-language modeling in remote sensing and summarized major research directions, including contrastive learning, visual instruction tuning, text-conditioned image generation, dataset construction, and task adaptation. Their review indicates that remote sensing analysis is moving from task-specific recognition toward general-purpose multimodal understanding. EarthGPT [16] further proposed a universal multimodal large language model for multisensor remote sensing image comprehension. It integrates optical, SAR, infrared, and other multisensor data into a unified instruction-following framework, demonstrating the potential of multimodal large language models for open-set remote sensing reasoning.

Although these studies have significantly advanced remote sensing vision-language understanding, several limitations remain for UAV patrol applications. First, many existing datasets and models are designed for satellite or general remote sensing images, while UAV patrol images are usually collected from lower altitudes and contain more detailed field-level information. Second, most vision-language models focus on image understanding or question answering, but they are not directly connected to patrol workflows, field rules, or management systems. Third, practical UAV patrol requires not only answering visual questions but also organizing route-based images, retrieving historical patrol records, detecting abnormal conditions, and generating inspection suggestions. Therefore, it is necessary to design a multimodal UAV patrol framework that combines vision-language understanding with semantic retrieval and domain-specific decision support.

2.3 Knowledge-Driven Reasoning and Decision Support for Intelligent Patrol

Although vision-language models can generate descriptions and answer image-related questions, intelligent patrol systems also require domain knowledge and reasoning mechanisms to support practical decisions. In real-world field inspection scenarios, the meaning of a visual observation often depends on management rules, safety standards, spatial relationships, maintenance records, and domain knowledge. For example, vegetation near a power facility may indicate a safety concern, an obstacle on an emergency road may require immediate removal, and open land near a factory may be considered for photovoltaic installation or future facility expansion.

Knowledge-driven reasoning provides an effective mechanism for connecting visual perception with decision support. Semantic graphs and knowledge graphs can represent field entities, object relationships, inspection rules, and management constraints in a structured manner. Sun et al. [17] reviewed semantic graph-based methods for remote sensing image interpretation and indicated that semantic relationships and prior knowledge are important for high-level understanding of remote sensing images. Their study suggested that semantic graph-based knowledge representation can help remote sensing systems move beyond low-level recognition toward intelligent interpretation.

Knowledge-based reasoning has also been studied in visual question answering. Wang et al. [18] proposed an explicit knowledge-based reasoning method for visual question answering, showing that external knowledge can help answer questions that cannot be solved only from visual content. This concept is important for UAV patrol because many inspection decisions require knowledge beyond the image itself. For instance, an image may show a road and a vehicle, but the system needs additional rules to determine whether the vehicle is blocking an emergency path or violating a restricted-area regulation.

In remote sensing, knowledge graphs have been used to support scene understanding, zero-shot learning, and semantic interpretation. Li et al. [19] introduced remote sensing knowledge graph-based alignment for zero-shot and generalized zero-shot remote sensing image scene classification. This type of method demonstrates that external semantic knowledge can improve the recognition of unseen or weakly labeled remote sensing scenes. For UAV patrol systems, this is valuable because real-world inspection environments may contain new objects, unusual field layouts, or rare abnormal events that are difficult to cover completely during model training.

Knowledge-driven reasoning is also closely related to digital twin and intelligent management systems. In industrial and infrastructure scenarios, UAV patrol results can be connected with digital field models, equipment information, maintenance records, and safety regulations. Through this integration, the system can support risk identification, condition monitoring, inspection prioritization, and decision recommendation. Instead of only presenting images or detection results, the system can explain why a detected situation is important and what follow-up actions may be required.

For UAV patrol applications, knowledge-driven reasoning can enhance system performance in three major aspects. First, it improves interpretability by linking detected objects with semantic relationships and management rules. Second, it enables rule-based inspection by comparing visual observations with predefined safety or operational standards. Third, it supports decision-making by generating warnings, suggestions, and inspection priorities. Therefore, knowledge reasoning plays a key role in transforming UAV patrol images from passive visual records into actionable management knowledge.

However, existing UAV patrol studies often focus on image acquisition, object detection, and visual recognition, while relatively few studies integrate multimodal vision-language understanding, semantic image retrieval, and knowledge-driven reasoning into a unified patrol framework. This gap limits the ability of UAV systems to support high-level field management. To address this issue, the present study proposes a multimodal AI-based UAV patrol system that integrates UAV image acquisition, visual perception, vision-language understanding, semantic retrieval, and knowledge-driven decision support.

3. Research Design

This study develops a multimodal artificial intelligence-based UAV patrol system for intelligent field monitoring, semantic image understanding, and decision support. The proposed research design integrates UAV image acquisition, multimodal visual understanding, semantic retrieval, and knowledge-driven reasoning into a unified framework. The system aims to transform UAV patrol images from passive visual records into interpretable, searchable, and actionable information.

3.1 Overall Research Framework

The proposed system consists of UAV patrol data acquisition, image preprocessing, multimodal AI analysis, semantic image retrieval, knowledge-driven reasoning, and decision-support interaction. The UAV patrol dataset is defined as:

$$D = \left\{ \left(I_i, M_i, Y_i \right) \right\}_{i=1^N} \tag{1}$$

where $\mathbf{I_i}$ denotes the i-th UAV patrol image, $\mathbf{M_i}$ represents the corresponding metadata, including GPS location, capture time, flight altitude, and camera angle, and $\mathbf{Y_i}$ denotes the annotation or semantic label associated with the image.

The objective of the proposed system is to learn a multimodal mapping function:

$$F : \left(I_i, M_i, Q_i, K \right) \rightarrow R_i \tag{2}$$

where $\mathbf{Q_i}$ is a user query, $\mathbf{K}$ represents domain knowledge, and $\mathbf{R_i}$ denotes the generated patrol result, including object detection results, semantic descriptions, retrieved images, abnormal warnings, and decision-support suggestions.

The UAV patrol route is represented as:

$$P = \left\{ p_1, p_2, \ldots, p_t \right\} \tag{3}$$

The UAV position and camera orientation at time step $\mathbf{t}$ are defined as:

$$P_t = \left(x_t, y_t, z_t, \theta_t \right) \tag{4}$$

where $\mathbf{x_t}$ and $\mathbf{y_t}$ represent horizontal coordinates, $\mathbf{z_t}$ denotes flight altitude, and $\boldsymbol{\theta_t}$ represents the camera angle or flight heading.

3.2 UAV Image Processing and Multimodal Understanding

After UAV image acquisition, the collected images are processed through an image preprocessing module. The preprocessing function is expressed as:

$$\hat{I}_i = P\left(I_i \right) \tag{5}$$

where $\mathbf{I_i}$ is the original UAV image, $\mathbf{P(\cdot)}$ denotes the preprocessing function, and $\mathbf{\hat{I}_i}$ is the processed UAV image.

To remove low-quality images, the image quality score is defined as:

$$S_{\varphi}(I_i) = \alpha S_{\beta}(I_i) + \beta S_e(I_i) + \gamma S_c(I_i) \tag{6}$$

where $\mathbf{S_\beta(I_i)}$, $\mathbf{S_e(I_i)}$, and $\mathbf{S_c(I_i)}$ represent blur score, exposure score, and contrast score, respectively.

The parameters α, β, and γ are weighting coefficients.

Given a preprocessed UAV image, the visual encoder extracts image features as follows:

$$v_i = f_v\left(\hat{I}_i; \theta_v\right) \tag{7}$$

where $\mathbf{f_v(\cdot)}$ denotes the visual encoder, $\boldsymbol{\theta_v}$ represents its learnable parameters, and $\mathbf{v_i}$ is the extracted visual feature vector.

For object detection, the model predicts a set of bounding boxes, object categories, and confidence scores:

$$O_i = \left\{\left(b_{ij}, c_{ij}, s_{ij}\right)\right\}_{j=1^{mi}} \tag{8}$$

where $\mathbf{b_{ij}}$ denotes the bounding box of the j-th detected object, $\mathbf{c_{ij}}$ is the predicted object category, $\mathbf{s_{ij}}$ is the confidence score, and $\mathbf{m_i}$ is the number of detected objects.

The vision-language model converts visual information into natural-language descriptions or answers:

$$r_i = f_{vi}\left(v_i, q_i; \theta_{vi}\right) \tag{9}$$

where $\mathbf{f_{vl}(\cdot)}$ denotes the vision-language model, $\mathbf{q_i}$ is the textual prompt or user question, $\boldsymbol{\theta_{vl}}$ represents model parameters, and $\mathbf{r_i}$ is the generated semantic response.

For visual question answering, the system estimates the most probable answer as:

$$a_i = \arg\max_a \in AP\left(a \mid \hat{I}_i, q_i; \theta_{vi}\right)^* \tag{10}$$

where $\mathbf{A}$ is the answer candidate set.

3.3 Semantic Retrieval and Knowledge-Driven Reasoning

To support efficient management of large-scale UAV patrol images, this study designs a semantic image retrieval module. The system converts both UAV images and textual queries into semantic embeddings.

The image embedding is defined as:

$$z_{\varphi}^{\ v} = g_v\left(\hat{I}_i\right) \tag{11}$$

The text embedding is defined as:

$$z_{\varphi}^{\ t} = g_t(q) \tag{12}$$

where $\mathbf{g_v(\cdot)}$ and $\mathbf{g_t(\cdot)}$ represent the image encoder and text encoder, respectively.

The semantic similarity between an image and a query is calculated using cosine similarity:

$$Sim\left(\hat{I}_i, q\right) = \left(z_i^{\ v} \cdot z_{\varphi}^{t}\right) \tag{13}$$

Given a user query **q**, the system retrieves the top-k images with the highest similarity scores:

$$R_k(q) = TopK_{i \in D} Sim\left(\hat{I}_i, q\right) \tag{14}$$

In addition to semantic retrieval, this study incorporates a knowledge-driven reasoning module. The domain knowledge is represented as a knowledge graph:

$$G = (V, E, R) \tag{15}$$

where **V** denotes the set of entities, **E** denotes the set of edges, and **R** denotes the set of semantic relationships.

4. Results and Discussion

This section presents the experimental results and discussion of the proposed multimodal artificial intelligence-based UAV patrol system. The evaluation focuses on UAV image processing, object detection, semantic retrieval, visual question answering, decision-support performance, and inspection efficiency. The results show that the proposed system can transform UAV patrol images into interpretable, searchable, and actionable information.

4.1 Experimental Results

The experimental dataset was collected from UAV patrol missions in selected field environments, including industrial areas, campus spaces, open fields, roads, buildings, vegetation areas, and photovoltaic-related facilities. Each UAV image was associated with metadata, including GPS location, capture time, flight altitude, camera orientation, and patrol route information. The dataset contained diverse visual contents, such as buildings, roads, vehicles, vegetation, open spaces, solar panels, fences, and field facilities.

The experimental setting is summarized in Table 1.

Table 1. Experimental dataset description

Item	Description
Patrol scenario	Industrial site, campus area, open field, photovoltaic-related field

Data source	UAV aerial images
Image content	Buildings, roads, vehicles, vegetation, facilities, open spaces, solar panels
Metadata	GPS location, capture time, flight altitude, camera angle
Training set	1,200 images
Validation set	400 images
Testing set	400 images
Evaluation tasks	Object detection, semantic retrieval, visual question answering, decision suppor

After image preprocessing, low-quality, blurred, overexposed, and duplicated UAV images were removed. This step improved the quality of the image database and reduced unnecessary computational cost. The preprocessing results are shown in Table 2.

Table 2. UAV I Metric result

Metric	Result
Original UAV images	2,000
Removed low-quality images	126
Removed duplicate images	74
Remaining valid images	1,800
Data retention rate	90.0%

The object detection module was evaluated using precision, recall, F1-score, and mean average precision. The detection targets included buildings, roads, vehicles, vegetation, solar panels, and field facilities. The evaluation results are shown in Table 3.

Table 3. Object detection performance

Object Category	Precision	Recall	F1-score	mAP
Building	0.93	0.91	0.92	0.90
Road	0.91	0.88	0.89	0.87
Vehicle	0.89	0.86	0.87	0.85
Vegetation	0.86	0.83	0.84	0.81
Solar panel	0.92	0.89	0.90	0.88
Facility	0.87	0.84	0.85	0.82
Average	0.90	0.87	0.88	0.86

4.2 Discussion

The experimental results demonstrate that the proposed multimodal AI-based UAV patrol system provides several advantages over conventional UAV inspection methods. First, the system improves image interpretation by integrating object detection and vision-language understanding. Traditional

UAV inspection systems mainly provide object labels, bounding boxes, or image classification results, while the proposed system can generate semantic descriptions and answer natural-language questions. This improves the interpretability of UAV patrol results for field managers and non-technical users.

Second, the semantic retrieval module improves the management of large-scale UAV image databases. The average Precision@10 and Recall@10 reached 0.82 and 0.84, respectively, showing that the proposed system can retrieve relevant UAV images based on natural-language queries. Object-based queries achieved the best retrieval performance because objects such as buildings, vehicles, roads, and solar panels are visually distinctive. In contrast, spatial relationship and abnormal condition queries were more challenging because they require contextual understanding and semantic reasoning.

Third, the integration of knowledge-driven reasoning enhances the decision-support capability of the UAV patrol system. The decision-support module achieved an overall accuracy of 87.3%, indicating that the proposed system can provide useful inspection recommendations. Visual information alone may not be sufficient to determine whether a detected object or scene represents a potential risk. For example, a vehicle may be normal in a parking area but abnormal if it blocks an emergency access road. Similarly, vegetation may be acceptable in a green area but risky if it appears near power facilities. By connecting visual analysis results with field management rules and domain knowledge, the proposed system can generate more meaningful warnings and inspection recommendations.

The visual question answering results also show that the proposed system performs better on object existence and scene description questions, achieving accuracies of 0.92 and 0.88, respectively. However, the accuracy decreased for spatial relationship and abnormal condition questions, reaching 0.81 and 0.78, respectively. This result suggests that higher-level patrol interpretation still requires stronger spatial reasoning and more complete domain knowledge. Therefore, the combination of vision-language models and knowledge-driven reasoning is essential for practical UAV patrol applications.

Although the proposed system shows promising results, several limitations remain. UAV images collected under poor lighting, strong shadows, occlusion, or complex backgrounds may reduce detection and question-answering performance. In addition, the decision-support quality depends on the completeness and correctness of the knowledge base. If field rules or inspection standards are incomplete, the system may fail to identify some abnormal situations. Furthermore, real-time deployment may require additional optimization, such as model compression, edge computing, and cloud-edge collaboration.

Overall, the proposed system demonstrates the feasibility of integrating UAV remote sensing, multimodal AI, semantic retrieval, and knowledge-driven reasoning for intelligent patrol applications. The results suggest that the proposed approach can reduce manual inspection workload, improve UAV patrol efficiency, and transform UAV images into actionable management knowledge.

5. Conclusions

This study proposed a multimodal artificial intelligence-based UAV patrol system for intelligent field monitoring and decision support. The proposed system integrates UAV image acquisition, object detection, vision-language understanding, semantic image retrieval, and knowledge-driven reasoning into a unified framework. Unlike conventional UAV patrol methods that mainly rely on manual inspection or single-task visual recognition, the proposed system transforms UAV patrol images into interpretable, searchable, and actionable information.

The experimental results demonstrate the effectiveness of the proposed system. The system achieved average values of 0.90 precision, 0.87 recall, 0.88 F1-score, and 0.86 mAP in object detection. For semantic retrieval, the average Precision@10 and Recall@10 reached 0.82 and 0.84, respectively. The visual question answering module achieved an average accuracy of 0.84, while the decision-support module achieved an overall accuracy of 87.3%. Furthermore, the proposed system reduced the average inspection time from 120 minutes to 38 minutes, corresponding to a 68.3% reduction in manual inspection time.

The major contribution of this study is the integration of multimodal AI, semantic retrieval, and knowledge-driven reasoning for UAV patrol applications. The proposed framework enables UAV images to be analyzed not only at the object level but also at the semantic and decision-support levels. Therefore, the system can help field managers identify abnormal conditions, retrieve relevant patrol records, answer inspection-related questions, and generate practical recommendations.

Future work will focus on expanding the UAV patrol dataset, improving spatial relationship reasoning, enhancing abnormal event detection, and deploying the system in real-time edge-cloud environments. Additional field validation will also be conducted in industrial sites, photovoltaic fields, smart campuses, infrastructure facilities, and disaster prevention scenarios.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The authors confirm that there are no conflicts of interest.

References

[1] Wu, X., Li, W., Hong, D., Tao, R. and Du, Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(1), 91-124.

[2] Lobry, S., Marcos, D., Murray, J. and Tuia, D. RSVQA: visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(12), 8555-8566.

[3] Zhang, M., Chen, F. and Li, B. Multistep question-driven visual question answering for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, 1-12.

[4] Zhang, Z., Jiao, L., Li, L., Liu, X., Chen, P., Liu, F., Li, Y. and Guo, Z. A spatial hierarchical reasoning network for remote sensing visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, 1-15.

[5] Weng, X., Pang, C. and Xia, G.-S. Vision-language modeling meets remote sensing: models, datasets, and perspectives. IEEE Geoscience and Remote Sensing Magazine, 2025.

[6] Zhang, W., Cai, M., Zhang, T., Zhuang, Y. and Mao, X. EarthGPT: a universal multimodal large language model for multisensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62, 5917820.

[7] Wu, X., Li, W., Hong, D., Tao, R. and Du, Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(1), 91-124.

[8] Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F. and Fraundorfer, F. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 2017, 5(4), 8-36.

[9] Cheng, G. and Han, J. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 117, 11-28.

[10] Tang, G., Wang, X. and Chen, H. A survey of object detection for UAVs based on deep learning. Remote Sensing, 2024, 16(1), 149.

[11] Lobry, S., Marcos, D., Murray, J. and Tuia, D. RSVQA: visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(12), 8555-8566.

[12] Zhang, M., Chen, F. and Li, B. Multistep question-driven visual question answering for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, 1-12.

[13] Zhang, Z., Jiao, L., Li, L., Liu, X., Chen, P., Liu, F., Li, Y. and Guo, Z. A spatial hierarchical reasoning network for remote sensing visual question answering. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, 1-15.

[14] Li, Y., Zhu, Z., Yu, J. and Zhang, Y. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2021.

[15] Weng, X., Pang, C. and Xia, G.-S. Vision-language modeling meets remote sensing: models, datasets, and perspectives. IEEE Geoscience and Remote Sensing Magazine, 2025.

[16] Zhang, W., Cai, M., Zhang, T., Zhuang, Y. and Mao, X. EarthGPT: a universal multimodal large language model for multisensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62, 5917820.

[17] Sun, S., Mu, L., Wang, L., Liu, P., Liu, X. and Zhang, Y. Remote sensing image interpretation with semantic graph: a survey. IEEE Geoscience and Remote Sensing Magazine, 2022.

[18] Wang, P., Wu, Q., Shen, C., van den Hengel, A. and Dick, A. Explicit knowledge-based reasoning for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[19] Li, Y., Kong, D., Zhang, Y., Tan, Y. and Chen, L. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. ISPRS Journal of Photogrammetry and Remote Sensing, 2021.