Journal of Electronic Research and Application, 2025, Volume 9, Issue 6
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Multi-Modal UAV Inspection of Photovoltaic Modules Using a YOLOv9-Based Fusion Network

Qing Yi[1], Jiayou Sun[2], Shanying Su[1], Houzhi Wei[1], Ke Wang[1], Zhihui Qi[1], Siyu Teng[3]*

[1]China Resources New Energy (Chibi) Co., Ltd., Chibi 437300, Hubei, China
[2]Waytous, Shenzhen 518000, Guangdong, China
[3]Shenzhen University, Shenzhen 518060, Guangdong, China

**Author to whom correspondence should be addressed.*

Abstract: The rapid expansion of photovoltaic (PV) power plants has created a pressing need for efficient and reliable operation and maintenance (O&M). Traditional manual inspection is slow, costly, and prone to error, motivating the use of unmanned aerial vehicles (UAVs) with infrared and visible cameras for automated monitoring. In this paper, we propose a YOLO-based multi-task framework for simultaneous PV defect detection and hazard-level classification. We constructed a dataset of 5,000 annotated UAV images from the Riyue PV power plant, covering ten defect categories and four severity levels (LV1–LV4). To support severity grading, the YOLO architecture was extended with a dual-task head and an ordinal regression scheme. The model was trained with a compound loss combining bounding-box regression, objectness, defect classification, and hazard-level supervision. Experimental evaluation on real UAV inspection data (224 strings, 30 ground-truth defects) shows that the proposed approach achieves mAP50 of 95.6%, recall of 92.7%, and severity classification accuracy of 90.8%. The system detects both minor anomalies (e.g., bird droppings, soiling) and critical faults (e.g., missing panels, disconnections) in real time at over 40 FPS, providing actionable insights for maintenance prioritization. These results demonstrate that YOLO-based UAV inspection offers a robust and scalable solution for intelligent PV O&M.

Keywords: Photovoltaic power plants; Unmanned aerial vehicles; Object detection; YOLOv9

Online publication: December 16, 2025

1. Introduction

The global deployment of photovoltaic (PV) systems has accelerated in recent years, making PV power plants major contributors to renewable energy. However, as the number of PV modules grows, inspection and maintenance have become more difficult. Manual inspection is labor-intensive, costly, and inefficient, especially for large-scale farms. This has driven the exploration of intelligent, autonomous inspection methods.

Unmanned aerial vehicles (UAVs) have developed rapidly and are widely used in industrial inspection. Their flexibility, low cost, and ability to capture visible-light and infrared (IR) images make them suitable for

autonomous PV inspection. Unlike ground-based inspection, UAVs can cover large areas quickly and provide multi-modal data for analysis.

PV module anomalies can be divided into two groups. Visible anomalies, observable from RGB imagery, include surface obstructions, missing or broken panels, and misalignments. Invisible anomalies, detectable only in infrared imagery, include abnormal temperature distributions, hot spots, and bypass diode failures. Since relying on a single modality leads to incomplete or inaccurate detection, RGB-IR fusion is needed. The schematic diagram of monitoring PV by UAVs is shown in **Figure 1**.

Figure 1. The schematic diagram of monitoring PV systems using UAVs

In this work, we constructed a dataset of 500 UAV-captured defective scenarios, covering both visible and thermal anomalies, and expanded it to 5,000 images through data augmentation. Based on this dataset, we developed a YOLOv9-based fusion detection network that jointly processes RGB and infrared images. The model fuses features from both modalities at multiple scales, enabling more robust defect localization than single-modality approaches.

The contributions of this paper are threefold:

(1) We construct a multi-modal UAV dataset of PV module defects, covering both visible and infrared anomaly types.

(2) We design a YOLOv9-based fusion detection network that leverages complementary RGB and IR features for accurate defect localization.

(3) We validate the effectiveness of the proposed approach on real inspection data, demonstrating improved detection robustness, accuracy, and real-time feasibility for large-scale PV power plants.

2. Related work

The task of intelligent PV module inspection intersects with multiple areas of computer vision research, namely 2D object detection, multi-modal fusion with infrared imagery (**Figure 2**), and UAV-based visual inspection. This section reviews the most relevant contributions in each domain.

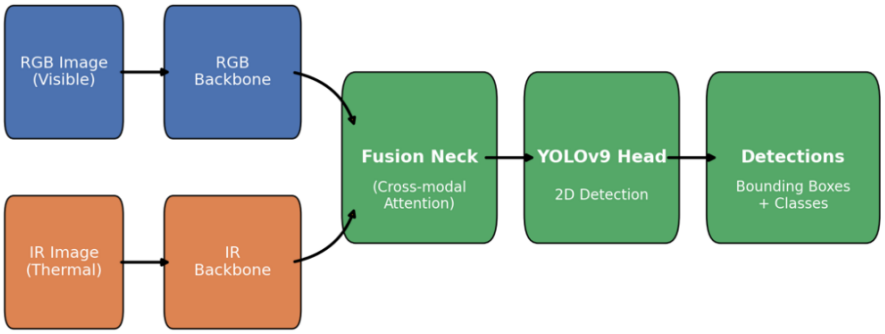

Figure 2. The fusion network for RGB and IR images

2.1. 2D object detection

2D object detection has been a core research problem in computer vision for decades. Traditional approaches relied on handcrafted features and sliding-window classifiers, which were often limited in complex real-world scenarios. The introduction of convolutional neural networks (CNNs) revolutionized detection performance, with models such as Faster R-CNN [1], SSD [2], and RetinaNet [3] enabling end-to-end training. Among them, the YOLO family of detectors has gained particular attention due to its real-time inference capability and competitive accuracy. The latest versions, such as YOLOv7–v9 [4–6], introduce anchor-free designs, decoupled heads, and improved feature aggregation, achieving strong trade-offs between speed and accuracy. These developments provide a solid foundation for applying 2D detection to UAV inspection tasks where efficiency and robustness are critical.

2.2. Infrared–RGB fusion detection

Infrared–RGB (IR–RGB) fusion detection has drawn more attention recently as the two modalities offer complementary information. RGB imagery provides fine texture and structural details but is sensitive to illumination changes, whereas infrared imagery shows thermal distributions and works well in low light. To leverage these complementary cues, many fusion-based object detection methods have been proposed. For instance, Liu *et al.* [7] first introduced multispectral deep networks for pedestrian detection, demonstrating that fusing IR with RGB improves performance in night-time scenes. Subsequent works, such as Takumi *et al.* [8] and CIAN [9], further enhanced robustness by introducing feature interaction mechanisms to better balance modality contributions. More recent detectors, such as MS-YOLO [10], extend the YOLO framework with cross-modal attention to adaptively combine RGB and IR features. Similarly, GFD-SSD [11] builds upon the SSD framework by introducing gated fusion units to learn the relative importance of different modalities. These advances show that IR–RGB fusion outperforms single-modality detection, especially under challenging conditions. However, most studies focus on surveillance and autonomous driving, with limited exploration of fusion-based detection in renewable energy inspection. This gap motivates our work on UAV-enabled PV monitoring using RGB–IR fusion.

2.3. UAV-based visual inspection

The application of UAVs for inspection has rapidly expanded in industries such as power transmission, wind energy, construction, and precision agriculture. UAV platforms, suitable for infrastructure monitoring, offer wide-area coverage, flexible flight planning, and large-scale multi-modal data acquisition. Recent studies explored autonomous flight path optimization, onboard visual analytics, and AI-based detector integration to

improve inspection efficiency. In the renewable energy sector, UAV-assisted inspection is vital for PV power plants to monitor thousands of modules for surface and thermal anomalies. Existing research used UAV-based thermographic analysis for hot spots and manual annotation of surface faults, but these methods have limitations in scalability, real-time processing, and full automation. To address these issues, this work uses UAV imagery with a fusion-based YOLOv9 detection framework to provide a robust and scalable pipeline for intelligent PV module inspection.

3. Proposed method

3.1. Overview

As illustrated in **Figure 2**, our detector takes as input a co-registered RGB image I_{rgb} and an infrared (IR) image I_{ir} captured by a UAV. The objective is to localize PV anomalies with 2D bounding boxes $b = (x, y, w, h)$, associated confidence scores, and defect categories. We adopt a YOLOv9-style one-stage framework and design a lightweight fusion neck that integrates complementary cues: RGB features emphasize surface appearance and structural integrity, while IR features highlight abnormal thermal responses.

3.2. Architecture

The overall architecture of the proposed detector (**Figure 2**) follows a three-stage design: dual backbones, a multi-modal fusion neck, and a YOLOv9 detection head.

Inputs: The system uses co-registered RGB and IR images from UAVs during PV inspection. The RGB input offers surface and structural cues, and the IR input shows thermal distributions. Both are normalized and resized to 1280 × 720.

Dual backbones: Each modality is processed by a lightweight YOLOv9 backbone. The RGB branch, for three-channel inputs, extracts texture- and structure-oriented features. The IR branch, for single-channel inputs, uses a parameter-adapted backbone with channel adaptors in the stem. Both output multi-scale feature maps at *P3, P4,* and *P5* with increasing channel dimensions.

Fusion neck: Features from the two backbones are integrated here. First, RGB and IR features at each scale are concatenated and passed through a 1×1 convolution. Then, a gated attention mechanism generates weights to balance the two modalities. The fused features are aggregated using an FPN–PAN structure to enhance semantic richness and spatial precision.

Detection head: The fused multi-scale features go to an anchor-free YOLOv9 detection head. Each scale has parallel branches for bounding box regression, objectness confidence, and defect classification. The regression branch predicts box offsets, and the classification branch outputs logits. It is fully convolutional and parameter-efficient.

Outputs: The detector produces bounding boxes $b^{\wedge}$, confidence scores, and class predictions, enabling automatic identification and localization of PV module anomalies in large-scale inspection data.

3.3. Loss function

The training objective adopts a compound loss function consistent with the YOLO detection paradigm, combining localization, objectness, and classification components. Formally, the total loss can be expressed as:

$L = \lambda_{box}L_{box}+\lambda_{obj}L_{obj}+\lambda_{cls}L_{cls}$,

where λ_{box}, λ_{obj}, and λ_{cls} are balancing coefficients.

Bounding box regression: The bounding box loss L_{box} uses the Complete IoU (CIoU) formulation, considering overlap, center-point distance, and aspect ratio. Compared with traditional $\ell 1$ or IoU losses, CIoU converges faster and optimizes more stably, crucial for small or elongated PV module defects.

Objectness prediction: The objectness branch predicts the probability of a candidate region having an object. Using a binary cross-entropy (BCE) loss following YOLO conventions, with the target value scaled by the predicted IoU between candidate and ground-truth boxes. This IoU-aware design enhances the correlation between localization quality and objectness confidence, improving non-maximum suppression (NMS) reliability.

Defect classification: To handle the imbalance between frequent minor defects (e.g., soiling) and rare critical anomalies (e.g., disconnections), focal loss is adopted L_{cls}. It down-weights easy negatives and focuses on hard, misclassified samples, improving sensitivity to underrepresented fault types.

Label assignment: A dynamic label assignment strategy like SimOTA achieves stable matching between predictions and ground-truth boxes. Unlike fixed IoU-threshold assignment, it adaptively selects positive samples based on spatial proximity and classification confidence, enabling effective supervision across different defect scales and densities.

In combination, these components offer a balanced objective for accurate localization, robust objectness estimation, and reliable defect categorization. This compound loss design suits the multimodal fusion framework for precise PV anomaly detection by jointly optimizing RGB and IR features.

3.4. Data processing

Before training, all RGB and IR image pairs are geometrically aligned for pixel-level correspondence, as small misalignments can reduce feature-fusion effectiveness. Each input is resized to 1280 × 720 for a balance between computational efficiency and defect-localization detail.

To enhance generalization, various data augmentation strategies are used. Global augmentations like Mosaic and MixUp combine multiple images to create synthetic samples, enriching object layouts and improving occlusion robustness. Local augmentations include random horizontal flipping, resizing, and Gaussian noise injection. For the RGB branch, photometric transformations are applied to simulate outdoor PV farm illumination.

Besides standard augmentations, modality dropout is incorporated. During training, either the RGB or IR stream is randomly masked with a 0.3 probability. This makes the detector maintain discriminative ability when one modality is degraded, and helps the fusion network balance complementary cues.

3.5. Training algorithm

The end-to-end training pipeline is summarized in Algorithm 1. In each iteration, paired RGB and IR images are augmented, passed through their backbones, fused at multiple scales, and processed by the YOLOv9 detection head. The model is optimized with the compound loss in Section 3.3. Parameters are updated using stochastic gradient descent with momentum (0.937), weight decay (5×10^{-4}), and a cosine learning rate schedule starting from 0.01. Early stopping based on validation $\text{mAP}_{50:95}$ is used to prevent overfitting.

4. Experimental results

4.1. Dataset description

We built a dataset of 3,000 annotated PV module images from UAV inspections, covering five anomaly types. It has 100 panel missing, 100 soiling, 700 internal faults, 1,500 bird droppings, 100 hot spots, and 500 panel

misalignment samples. Internal faults and hot spots are only visible in IR images, while the other three can be identified from RGB images. This shows the complementary strengths of RGB and IR modalities, emphasizing the need for a fusion-based detection framework.

4.2. Detection performance

The YOLOv9-based fusion detector was trained on a dataset of 3,000 annotated images, split into 2,400 for training and 600 for testing. On the held-out test set, the model achieved an overall average precision (AP) of 98.2%, confirming the effectiveness of RGB–IR fusion for robust anomaly detection.

Table 1 reports per-class AP results for six representative categories. The detector performed well, with all categories having AP above 96%. RGB-visible anomalies (panel missing, bird droppings) had near-perfect detection rates, and IR-specific anomalies (internal faults, hot spots) also achieved high accuracy, showing the benefit of multi-modal fusion. The detection results are shown in **Figures 3** and **4** (RGB on the left, IR on the right). RGB can accurately localize anomalies such as panel missing, while internal faults and hot spots are hard to identify from RGB alone, but are evident in IR. This comparison emphasizes the need for RGB–IR fusion in PV anomaly detection. The proposed detector can accurately identify both visible anomalies and invisible thermal defects. Infrared information significantly improves detection robustness for non-distinguishable categories in RGB images.

Table 1. Per-class AP (%) of the YOLOv9 RGB–IR fusion detector on the test set

Defect category	AP (%)
Panel missing	99.1
Dirty	97.6
Panel fault (IR high in all)	96.8
Bird droppings	99.3
Hot spot (IR high in part)	96.5
Panel misalignment	98.9
Overall	98.2

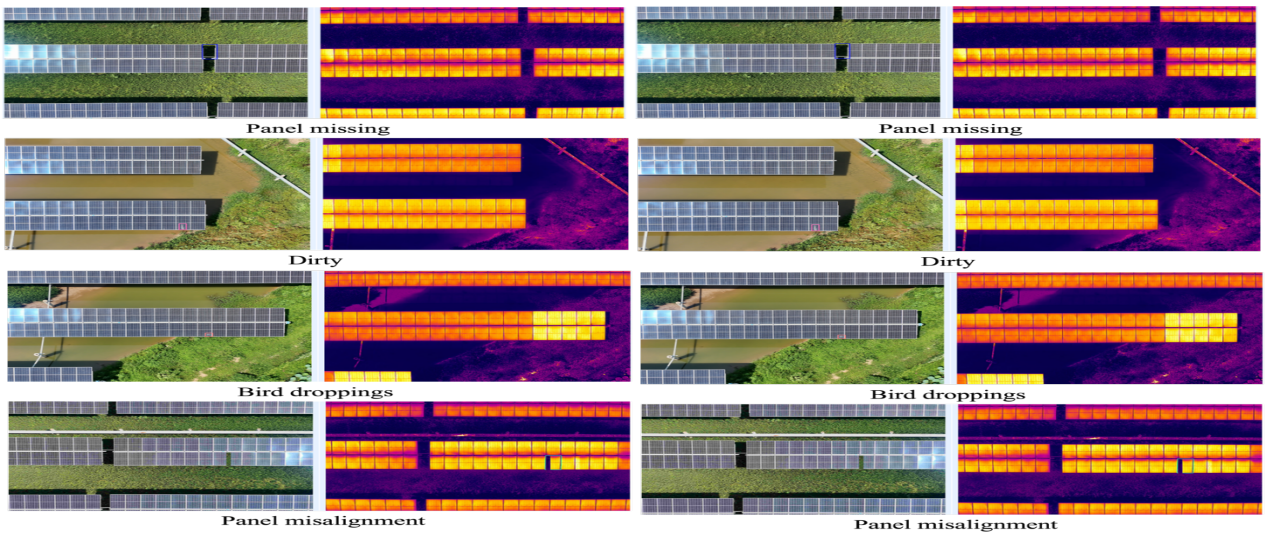

Figure 3. The detection results on the panel missing, dirty, bird droppings, and panel misalignment

4.3. Ablation study

To verify the effectiveness of the proposed dual-backbone fusion design, we conducted ablation experiments by comparing different input modalities and fusion strategies. The results are summarized in **Table 2**.

RGB-only: Training the detector with RGB images alone achieved strong performance on visible anomalies such as panel missing and bird droppings, but failed to identify thermal-related defects (internal faults and hot spots), leading to a drop in overall AP.

IR-only: Using only infrared inputs allowed the model to capture thermal anomalies but resulted in poor recognition of surface-level issues such as soiling or misalignment. This indicates that IR alone is insufficient for comprehensive PV inspection.

Naive concatenation: A baseline where RGB and IR channels were simply concatenated at the input and processed by a single backbone showed limited improvement over single modalities. Without dedicated branches, the backbone failed to balance heterogeneous features, resulting in sub-optimal fusion.

Proposed fusion network: Our dual-backbone design with gated fusion in the neck achieved the best performance, significantly improving detection accuracy across all categories. This validates the necessity of separating feature extraction per modality before fusion.

The comparison demonstrates that multi-modal fusion is essential for reliable PV anomaly detection, and that a dual-backbone design is markedly superior to naive input concatenation. These results highlight the complementary nature of RGB and IR features and the importance of a carefully designed fusion mechanism.

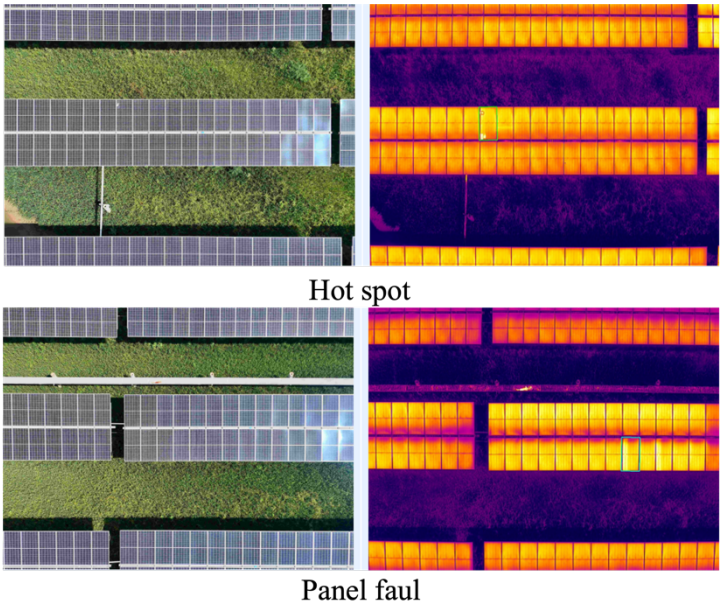

Figure 4. The detection results on the panel fault and hot spot

Table 2. Ablation results on the test set (AP, %)

Method	Overall AP (%)
RGB-only	91.7
IR-only	88.9
Naive concatenation	94.3
Proposed fusion (ours)	98.24

4.4. Discussion

The experimental results confirm the effectiveness of the proposed YOLOv9-based fusion framework for UAV-assisted PV inspection. The 98.24% overall AP on the test set indicates the model can reliably detect diverse anomaly types. Per-class evaluation shows high-precision recognition of RGB-visible and IR-only anomalies, highlighting the complementarity of the two modalities.

There is a difference in feature saliency between RGB and IR imagery. RGB visual anomalies are obvious and easy to detect, while IR anomalies are subtle and hard to capture, which is why the IR-only baseline performed worse than the RGB-only one.

The ablation study emphasizes the need for multi-modal fusion. The RGB-only model is good for surface-level defects but fails on thermal anomalies, and the IR-only model is effective for thermal faults but poor on visible categories. A simple concatenation strategy is sub-optimal, while the dual-backbone fusion design outperforms all baselines.

Qualitative analysis shows the proposed model offers more stable localization under challenging conditions, and the modality dropout strategy during training improves robustness.

Overall, effective use of RGB and IR imagery is crucial for large-scale PV inspection. The proposed architecture achieves state-of-the-art performance and practical robustness for real-world UAV deployment. Future work may explore lightweight variants or self-supervised pretraining.

5. Conclusion

This paper presented a YOLOv9-based RGB–IR fusion detector for UAV-assisted PV inspection. To tackle the challenges of monitoring large-scale PV plants, we designed a dual-backbone architecture with a gated fusion neck to integrate surface and thermal cues. The RGB branch captures structural and appearance anomalies, and the IR branch highlights temperature-related faults. Combining these modalities, the method offers a unified solution for detecting visible and invisible anomalies hard to capture with a single sensor.

Experiments on a UAV dataset showed the approach's effectiveness. The model achieved an overall average precision of 98.2% across six defect types, outperforming RGB-only, IR-only, and naive concatenation baselines. These results validate modality fusion and the superiority of the dual-backbone design in balancing heterogeneous features. The findings suggest that reliable PV anomaly detection needs to use the complementary strengths of RGB and IR imagery.

Funding

This work was supported in part by the Joint Innovation Program of Guangdong under Grant 2023A0505020003.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Girshick R, 2015, Fast R-CNN, Proceedings of the IEEE International Conference on Computer Vision, 1440–1448.

[2] Liu W, Anguelov D, Erhan D, et al., 2016, SSD: Single Shot Multibox Detector, European Conference on Computer Vision, 21–37, Springer.

[3] Lin TY, Goyal P, Girshick R, et al., 2017, Focal Loss for Dense Object Detection, Proceedings of the IEEE International Conference on Computer Vision, 2980–2988.

[4] Wang CY, Bochkovskiy A, Liao HYM, 2023, YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7464–7475.

[5] Varghese R, Sambath M, 2024, YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness, 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 1–6, IEEE.

[6] Wang CY, Yeh IH, Liao HYM, 2024, YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information, European Conference on Computer Vision, 1–21, Springer.

[7] Liu J, Zhang S, Wang S, et al., 2016, Multispectral Deep Neural Networks for Pedestrian Detection. arXiv. https://doi.org/10.48550/arXiv.1611.02644

[8] Takumi K, Watanabe K, Ha Q, et al., 2017, Multispectral Object Detection for Autonomous Vehicles, Proceedings of the Thematic Workshops of ACM Multimedia 2017, 35–43.

[9] Zhang L, Liu Z, Zhang S, et al., 2019, Cross-Modality Interactive Attention Network for Multispectral Pedestrian Detection. Information Fusion, 50: 20–29.

[10] Xie Y, Zhang L, Yu X, et al., 2023, YOLO-MS: Multispectral Object Detection Via Feature Interaction and Self-Attention Guided Fusion. IEEE Transactions on Cognitive and Developmental Systems, 15(4): 2132–2143.

[11] Zheng Y, Izzat IH, Ziaee S, 2019, GFD-SSD: Gated Fusion Double SSD for Multispectral Pedestrian Detection. arXiv. https://doi.org/10.48550/arXiv.1903.06999