A Comparison of CNN, Transformer, and Open-Vocabulary Architectures for Real-Time Photovoltaic Defect Detection Using UAV Thermal Imagery

Aissam Salah[1], Mouad Jabrane[2], Imane Sebari[1]

[1]Department of Photogrammetry and Cartography, School of Geomatics and Surveying Engineering, IAV Hassan II, Rabat, Morocco
aissamsalah89@gmail.com i.sebari@iav.ac.ma
[2]Research Unit of Geospatial Technologies for a Smart Decision, IAV Hassan II, Rabat 10101, Morocco
jmouad25@gmail.com

Keywords: Defect Detection, Photovoltaic Inspection, UAV Imagery, Object Detection, Open-Vocabulary Detection, Transformers

Abstract

Real-time defect detection in solar farms is critical for profitability and safety. This paper compares state-of-the-art (SOTA) object detection architectures for deployment on edge computing platforms for the purpose of thermal PV defect detection using UAV imagery. We systematically evaluated Closed-Set (YOLOv10, YOLOv12, RT-DETR, RF-DETR) and Open-Vocabulary (YOLO-World, OWL-ViT) models on a public thermal dataset. Our results highlight two leading candidates. The transformer-based RF-DETR sets a new accuracy record at 82.6% mAP@0.50, driven by its self-supervised backbone, but its inference speed is low (12.6 FPS). Conversely, the CNN-based YOLO-World integrates language semantics to reach a competitive 78.1% mAP@0.50 while operating at a real-time speed of 31.3 FPS. We conclude that both RF-DETR and YOLO-World are promising for embedded thermal fault detection. The final selection will depend on on-platform inference performance.

1. Introduction

Photovoltaic (PV) systems are experiencing exponential growth worldwide, driven by increasing energy demand and environmental imperatives (Akram et al., 2019). Although photovoltaic modules are designed for a long service life, their continuous exposure to environmental and operational stresses makes them susceptible to various forms of degradation that can affect their performance and reliability. Faced with the diversity of photovoltaic defects, several inspection methods have been developed. Among the imagery techniques, infrared thermography (IRT) is the most used for large-scale drone inspections (Barraz et al., 2025; De Oliveira et al., 2020).

Manual inspection methods are becoming increasingly impractical in the large solar farms due to the time consumption and the possibility of human errors. This context has driven the adoption of Deep Learning (DL) for PV defect identification (Masita et al., 2025). Current literature on UAV thermal imagery analysis focuses predominantly on Convolutional Neural Networks (CNNs) for real-time applications. To date, the adoption of other state-of-the-art architectures like Vision Transformers (ViT) remains less explored in PV defect detection (Zefri et al., 2023). Furthermore, there has been no exploration of Open Vocabulary Detection (OVD) as introduced by Vision-Language Models (VLMs), since current classification of PV defects is limited to predefined classes seen during model training (Al Mahdi et al., 2024; Hijjawi et al., 2023).

The aim of this work is to present one of the first systematic investigations of OVD for thermal-based PV inspection. We establish a new performance benchmark using the latest 2024 and 2025 architectures. The study evaluates cutting-edge models like YOLOv12 and RF-DETR for the first time in this field. A new accuracy record is established on the public "ThermoSolar-PV" dataset. The transformer-based RF-DETR achieves a record 82.6% mAP@0.50. We identify an optimal solution for real-time drone deployment. YOLO-World is validated as the best balance of open-vocabulary flexibility and 31.3 FPS speed.

2. Related Work

In recent years, the IRT-based PV defect detection was treated as a classification problem. For instance, the hotspots detection using traditional ML algorithm the Support Vector Model (SVM) to modifying VGG16 architecture to distinguish between three classes (healthy, hotspots and substrings) using a dataset acquired by an UAV and IR cameras (Hijjawi et al., 2023). However, real time object detection was dominated by the family of "You Only Look Once", the original YOLOv5 has achieved a mean precision of 83.86% in a drone thermal images dataset, this accuracy will increase by 4.4% after the introduction of a ShuffleNetV2 backbone and an attention mechanism in the original YOLO framework (Masita et al., 2025).

The object detection field expands to a more complete use of attention based models which can better determine global dependencies in the image. DETR (Detection Transformer) use a CNN to generate image features that are added to the position encoding and processed by the Transformer encoder. The outputs are then decoded with a learnable object queries to bounding boxes and category labels avoiding the Non Suppression Maximum that categorizes YOLO models (Li et al., 2024). But due to its high computational cost and quadratic complexity of a global attention mechanism more edge adapted versions are developed to achieve real time inference speed.

Despite the high accuracy of the DETR variants, they remain closed-set. They are restricted to a fixed number of classes defined during the training. In real-world solar farms, defects are unpredictable and constantly evolving. The models are incapable of detecting new, hybrid or rare PV defects. Open-Vocabulary Detection (OVD) provides a solution to this rigidity. OVD uses Vision-Language Models (VLMs) to link visual features with text descriptions. This replaces fixed classification heads with a dynamic comparison in a shared vector space. This approach offers unprecedented flexibility and reduces reliance on exhaustive annotations, as the model can generalize to new

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-261-2026 | © Author(s) 2026. CC BY 4.0 License.

classes based on its pre-existing linguistic knowledge (Minderer et al., 2022).

3. Method

To provide a systematic comparison of deep learning architectures for PV defect detection, The methodology is divided into three steps (Figure 1): data preprocessing that includes annotations' conversion and semantic rationalization of classes for the OVD models, a finetuning of the evaluated architectures, and performance evaluation.

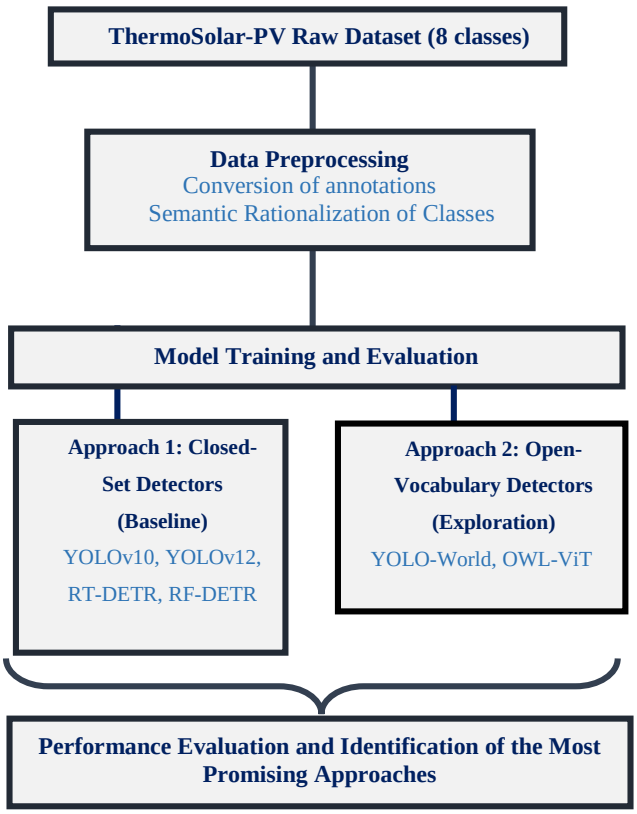


Figure 1. Flowchart of the General Methodology.

3.1 Dataset and Preprocessing

The study uses the "ThermoSolar-PV" public dataset, which contains 2,723 thermal UAV images with 7,772 labeled objects (Darabi, 2025). The images were resized to a 640×640 resolution with grayscale normalization and data augmentation for generalization. The dataset was originally partitioned into three predefined sets: training (6,924 images), validation (400 images), and test (15 images) and came with a baseline performance using the YOLOv9 model that achieved 78% mean precision (map@0.5).

The preprocessing has included the conversion of the annotations from the original YOLO format to the COCO JSON format that the transformer detection architectures are trained on. Second, for experiments involving OVD models, a semantic rationalization of the eight original defect classes was performed. Class pairs distinguished instance count (e.g., "SingleHotSpot", "MultiHotSpot") were consolidated into five distinct semantic categories: Hot Spot Defect, Diode Anomaly, Bypassed Cell Defect, String Open Circuit, and String Reversed Polarity. This step was implemented to provide language-vision models with clearer, non-redundant class definitions.

3.2 Evaluated Architectures

A selection of state-of-the-art object detection models was evaluated. All models were initialized with weights pre-trained on the MS COCO dataset and fine-tuned for 50 epochs on the dataset.

3.2.1 CNN-Based Architectures

CNN-based architectures were selected to establish a high-speed performance baseline. They remain the industry standard for real-time UAV-based inspection tasks. The selection focused on the latest YOLO iterations (v10 and v12) to represent the 2024-2025 state-of-the-art (SOTA). These models were chosen because they specifically aim to break the traditional performance-efficiency boundaries of previous generations like YOLOv8 and YOLOv9.

YOLOv10 was chosen for its breakthrough in eliminating the post-processing bottleneck while maintaining superior accuracy-latency trade-offs. It introduces a dual assignment strategy to achieve NMS-free inference, allowing for truly end-to-end detection. It employs a lightweight classification head and rank-guided block design to enhance model capability under low computational cost. In SOTA benchmarks, YOLOv10-S demonstrates significantly higher efficiency than RT-DETR-R18 while achieving comparable accuracy. The "s" (small) version was selected to respect the VRAM limits of edge computing platforms. With approximately 7.2 million parameters, it is optimized to provide high-speed processing within limited memory constraints (Wang et al., 2024).

YOLOv12 was selected as the 2025 state-of-the-art to evaluate the success of an attention-centric design in matching CNN speeds. It was chosen because it outperforms previous advanced models (YOLOv10 and YOLOv11) in mAP accuracy across all scales. The model integrates a linear area attention (A2) mechanism that maintains a large receptive field while drastically reducing complexity. It utilizes Residual Efficient Layer Aggregation Networks (R-ELAN) to solve optimization challenges and stabilize feature aggregation. Architectural heatmaps prove that YOLOv12 provides clearer object contours and more precise foreground activation compared to YOLOv10. A Position Perceiver (7x7 separable convolution) is used to help the model capture global dependencies more effectively than traditional CNNs. The "s" (small) variant was chosen to maintain comparability with the YOLOv10 baseline. It contains 9.3 million parameters, successfully balancing advanced Transformer-like reasoning with the lightness required for real-time drone deployment (Tian et al., 2025).

3.2.2 Transformer-Based Architectures

Transformer-based models were selected to address the inherent limitations of CNNs in modelling global dependencies. Unlike CNNs, which focus on local receptive fields, Transformers use self-attention to capture long-range contextual relationships within the thermal imagery. This family was included to evaluate the "end-to-end" detection paradigm, which simplifies the pipeline by removing hand-crafted components like anchor boxes and Non-Maximum Suppression (NMS). The selection represents the 2024-2025 shift in SOTA where Detection

Transformers (DETR) began competing with YOLOs in real-time scenarios.

RT-DETR was selected as the first real-time end-to-end Transformer-based detector to successfully outperform traditional YOLOs in both speed and accuracy. It was chosen for its Efficient Hybrid Encoder, which decouples intra-scale interaction and cross-scale fusion to solve the computational cost of standard Transformers. The model utilizes Attention-based Intra-scale Feature Interaction (AIFI) to limit expensive self-attention only to high-level features where it is most effective. It introduces Uncertainty-minimal Query Selection, which selects high-quality initial queries for the decoder by explicitly optimizing for localization confidence. This architecture is uniquely scalable, allowing the number of decoder layers to be adjusted to balance speed and precision without retraining. The "Large" (L) version (ResNet-50 backbone) was chosen to evaluate the potential of a deeper Transformer structure in complex thermal environments. This variant provides a robust benchmark that rivals the most advanced YOLOs while maintaining a respectable speed profile for high-end edge platforms (Zhao et al., 2023).

RF-DETR was selected for its record-breaking precision. It was chosen specifically for its DINOv2 backbone, a Vision Transformer (ViT) pre-trained via self-supervision on massive datasets. This backbone allows the model to learn highly generalizable visual representations, making it exceptionally effective at identifying subtle thermal anomalies with limited specialized data. The architecture utilizes deformable attention mechanisms to focus computation on relevant spatial zones, improving detection under heavy occlusion or camouflage. For real time application, we selected it based on NVIDIA experiments showing massive optimization potential. Documentation indicates that post-training via NVIDIA TensorRT can reduce latency to 6.0ms (~166 FPS) on compatible hardware, proving its long-term viability for real-time drone platforms. The "Base" (B) variant (approximately 29M parameters) was selected as the optimal version, offering a superior balance between moderate parameter count and state-of-the-art performance (Sapkota et al., 2025).

3.2.3 Vision-Language Models (VLMs)

Vision-Language Models (VLMs) were selected to overcome the rigid "closed-set" limitation of CNNs and Transformers. These models were included to evaluate Open-Vocabulary Detection (OVD), allowing for the detection of diverse and evolving PV defects through text prompts. By leveraging semantic knowledge from large-scale image-text pre-training, these architectures can generalize to rare or hybrid anomalies without requiring exhaustive new annotations.

YOLO-World was selected as the primary OVD candidate for its industry-leading balance of open-vocabulary flexibility and real-time speed. It was chosen for its innovative "prompt-then-detect" strategy, which decouples heavy language reasoning from the fast detection process. The model utilizes a Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN) to efficiently fuse visual and linguistic features. A critical advantage of this architecture is its ability to re-parameterize text embeddings directly into the model's weights after training. This allows the heavy text encoder to be completely removed during deployment, enabling the model to operate at the speed of a standard YOLOv8. Benchmarks show that YOLO-World-L achieves comparable accuracy to Grounding DINO while being over 30 times faster.The "Small" (S) version (approximately 25M parameters in its V2 variant) was chosen to prove that OVD capabilities can be successfully integrated into lightweight edge-ready backbones (Cheng et al., 2024).

OWL-ViT was selected as a pioneer in Vision Transformer-based OVD, using a simple architecture with minimal modifications to standard ViTs. It was chosen for its modularity, as it can perform both text-conditioned and one-shot image-conditioned detection by matching object embeddings in a shared space. The model leverages contrastive pre-training from CLIP to align visual representations with rich semantic descriptions. While standard OWL-ViT is computationally heavy, we included it based on the NVIDIA NanoOWL optimization project. NVIDIA experiments demonstrate that by using TensorRT and FP16 quantization, OWL-ViT can be accelerated to reach 25–95 FPS on Jetson Orin platforms. This optimization path makes a traditionally slow Transformer model viable for real-time robotic and drone applications. The ViT-B/32 variant was selected as the optimal version for this study because its inference compute is comparable to a ResNet-50 (Carion et al., 2020).

3.3 Experimental Setup and Metrics

Training and inference were executed on an NVIDIA GeForce RTX 4050 Laptop GPU. The software environment was built on Python 3.9 using the PyTorch framework, the libraries that managed the training and assured the evaluation are Ultralytics for YOLO models and Hugging Face Transformers.

Model performance was evaluated using the main metrics for object detection. The mean Average Precision with Intersection over Union of 0.5 (mAP@0.50) is used to evaluate the model's ability to correctly classify defects and the mean Average Precision for IoU thresholds from 0.5 to 0.95 (mAP@0.50:0.95) to evaluate the localization accuracy. The suitability of each model for real-time deployment was measured by its Inference Speed in Frames Per Second (FPS).

4. Results

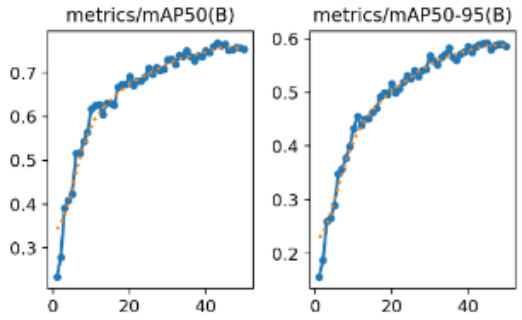
This section presents the results of the benchmark. It tackles firstly the training evolution during the 50 epochs and the convergence of the different examined models. Secondly a comparison of each model's best checkpoint performance on the validation set according to the predefined metrics. Finally, the generalization of models on the test set and visualisation of a sample's predictions compared to the ground truth.

4.1 Training Dynamics

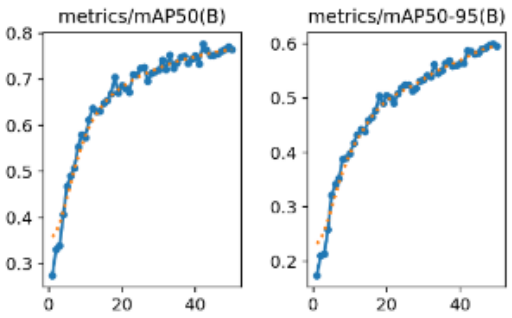
The metrics were monitored for each model over 50 epochs on the validation set. YOLOv10s and YOLOv12s, produced stable convergence. The loss curves (Figure 2) show a constant and regular decrease throughout the 50 epochs. The accuracy metrics show rapid progress during the first 15-20 epochs, before fluctuating around a high plateau without significant signs of overfitting. YOLO-World displayed a similarly stable and rapid convergence pattern.

In contrast, the Transformer-based architectures showed more varied training dynamics. RT-DETR-L demonstrated greater volatility in its validation metrics, with more more pronounced peaks from one epoch to another. The model was more sensitive to variations between validation batches.

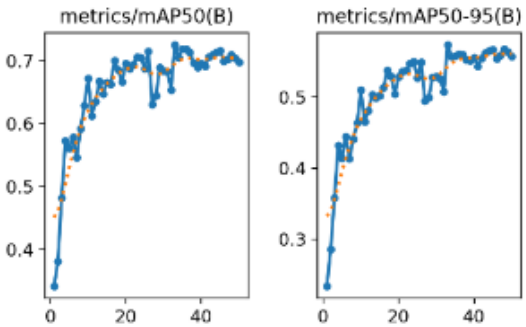
RF-DETR-B accordingly to its literature, achieving a fast convergence within the first 10-15 epochs. The performance of its Exponential Moving Average (EMA) checkpoint surpassed the base model, highlighting the effectiveness of this regularization technique for stabilizing the fine-tuning process.



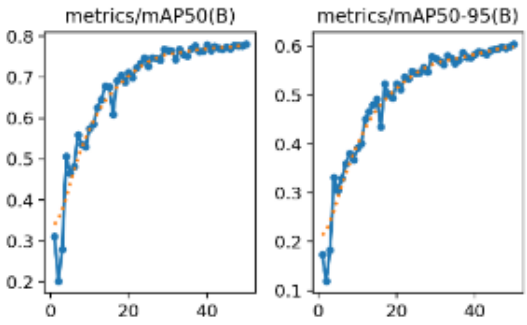
YOLO v10



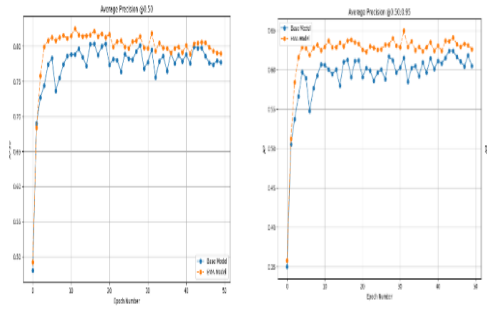
YOLO v12



RT-DETR



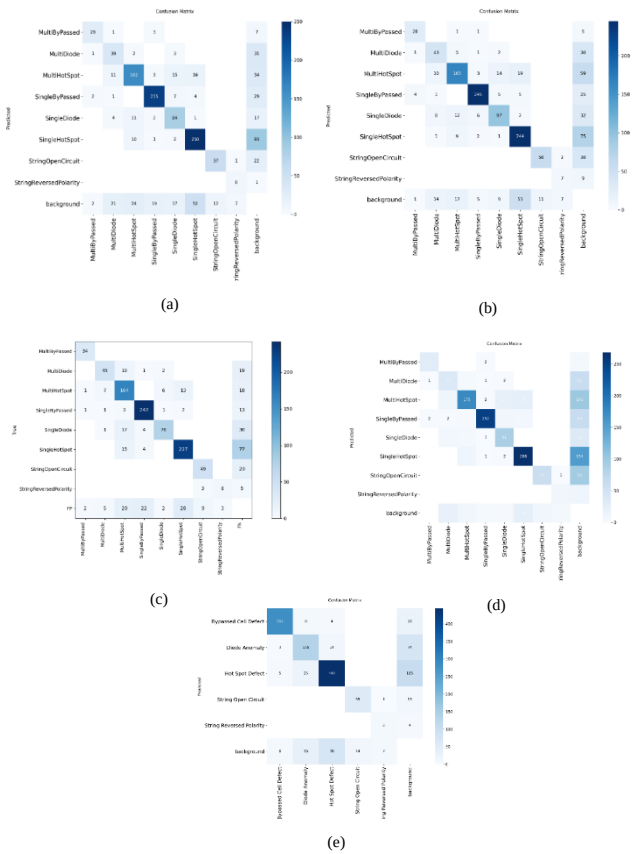
YOLO World (5 cls)



RF-DETR

Figure 2. Progression of validation mAP@0.50 (left) and mAP@0.50:0.95 (right) over 50 training epochs.

The confusion matrix of each model (Figure 3) presents the accuracy of detection for each class for the validation set. All models show various errors in the distinction between the Multi and Single instances of the same defect and the hotspot and diode anomaly classes due to their similar thermal signature. RF-DETR and YOLOv12 show fewer of this errors. This indicates that the rationalization has a double effect that even the closed set models could benefit from. After the visualisation of samples of the models' annotations, the background detections may conclude undetected defects by the authors of the dataset.



(a) (b) (c) (d) (e)

Figure 3. Confusion matrix for models' best checkpoint on the validation set. (a) YOLO v10 (b) YOLO v12 (c) RT-DETR (d) RF DETR (e) YOLO World

4.2 Final Performance Evaluation

Table 1 summarizes the performance results of the best checkpoint for each model. It presents the accuracy, inference speed and the number of parameters.

RF-DETR-B outperforms all models and sets a new state-of-the-art for precision in this task, achieving a mean Average Precision (mAP@0.50) of 82.6% and a mAP@0.50:0.95 of 65.0%. This performance is explained by the advanced feature extraction capabilities of its DINOv2 backbone.

In terms of inference speed, the YOLO architectures remains superior. YOLOv10s and YOLOv12s recorded the highest inference speeds at 37.10 FPS and 37.23 FPS, respectively. RT-DETR-L, while more computationally complex, attained a speed of 30.08 FPS. The RF-DETR-B model, however, recorded a significantly lower speed of 12.59 FPS, rendering it incapable of real time application prior to any post-training optimization.The OVD models produced two outcomes. YOLO-World, achieved an mAP@0.50 of 78.1%, while operating at a real-time speed of 31.34 FPS. In contrast, the fine-tuning of OWL-ViT model proved challenging. Despite multiple attempts with different training configurations, its performance remained low, with a final mAP@0.50 of 19.7%. and a high recall of 68.0%. This indicates that while the model could successfully localize thermal anomalies, it failed to classify them correctly.

Model	mAP@0.50	mAP@0.50:0.95	FPS	Parameters (M)
YOLOv10s	0.754	0.586	37.1	7.2
YOLOv12s	0.771	0.602	37.2	9.3
RT-DETR-L	0.725	0.573	30.1	42.0
RF-DETR-B	0.826	0.650	12.6	29.0
YOLO-World	0.781	0.599	31.3	25.8

Table 1. Final performance and complexity comparison of the evaluated models, captured at the best checkpoint on the validation set.

4.3 Qualitative and Generalization Analysis

A final evaluation was conducted on a test set. This set of 15 images was not used at any point during training. For each architecture, the mAP scores obtained on the test set were higher than those observed on the validation set, confirming that no significant overfitting had occurred and that the models could generalize to unseen data.

Visual analysis of the models' predictions on the test set confirms the metrics' evaluation. As illustrated in sample predictions in Figure 4, RF-DETR-B consistently produces precise bounding boxes and correctly classifies even subtle or challenging defect instances. YOLOv12 likewise demonstrates robust and accurate detections. The predictions from YOLOv10 and RT-DETR, while generally effective, show a higher tendency for the types of classification errors observed in the confusion matrices, such as confusing single and multi-instance defects and wrong background detections. The performance of YOLO-World is visually very strong, with a high degree of classification accuracy.

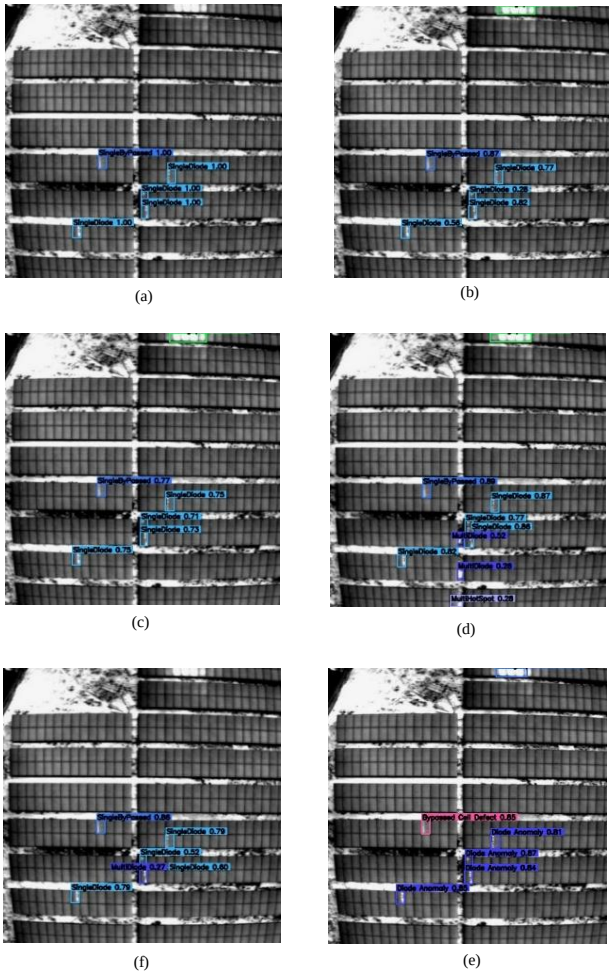


Figure 4. Qualitative comparison of detection results on a sample image from the test set. (a) Ground Truth, (b) YOLOv10s, (c) YOLOv12s, (d) RT-DETR-L, (e) RF-DETR-B, (f) YOLO-World

5. Discussion

The YOLO10 and, even more so, RT-DETR models present the most significant classification difficulties. They are prone to confusion errors between the Single and Multi-variants of the same defect, as well as between the "Diode Anomaly" and "HotSpot" classes. Furthermore, these two architectures have a greater tendency to miss actual defects, incorrectly assigning them to the background class.

RF-DETR and, to a lesser extent, YOLOv12 models prove to be more robust. Remarkably, several detections classified as background (false positives) by these models turn out, upon visual inspection, to be either actual defects not annotated in the ground truth or anomalies of very low thermal intensity. This

ability to identify subtle, unlabelled defects demonstrates an excellent degree of generalization and very high detection sensitivity.

YOLO-World, although architecturally based on an older backbone (YOLOv8), manages, thanks to the integration of semantic capabilities, to achieve a performance level (78.1% mAP@0.50) that rivals YOLOv12s. This result demonstrates that enrichment via a Vision-Language model can, on its own, close the performance gap with more modern, purely visual architectures.

The inability of the OWL-ViT model to achieve competitive performance can be attributed to a combination of two factors. The OWL-ViT architecture relies on a Vision Transformer pre-trained on billions of real-world (RGB) images. Its backbone has therefore learned to extract visual features such as colors, complex textures, common object shapes, and so on. Thermal imaging, on the other hand, is a radically different domain, where the relevant information resides in subtle temperature gradients and heat patterns that bear little resemblance to the characteristics of natural images.

Although it is a VLM, OWL-ViT is relatively a lightweight architecture for open-vocabulary detection. Newer and more powerful models, such as Grounding DINO, incorporate much deeper cross-modality fusion mechanisms, with dedicated encoders and cross-attention layers that allow for finer alignment between visual and textual representations. It is likely that the simpler architecture of OWL-ViT, where text-image interaction occurs primarily at the classification head level, is insufficient for the complexity of our fine-tuning task.

This work opens several promising perspectives of our research to overcome the identified limitations. The first focuses on exploring knowledge distillation learning from a powerful VLM model. The most promising way to obtain an OVD model that is both extremely accurate and efficient for the Edge would be to use a state-of-the-art teacher model, such as Grounding DINO, to teach a more compact and lightweight student model.

The second concerns improving the dataset. The performance and flexibility of OVD models are inherently limited by the size and semantic richness of the training data. Creating a much larger dataset, containing thousands of examples for each type of defect and, crucially, enriched with varied and contextual textual descriptions, would be a major contribution to the field. Such a dataset would not only improve the accuracy of current models but also truly unlock the zero-shot potential of VLMs, enabling them to identify rare or previously uncatalogued anomalies.

6. Conclusion

At the end of this study, we investigated object detection architectures for the automated inspection of thermal anomalies in photovoltaic panels. The objective was to identify, evaluate, and compare the state of the art models for real-time applications. The approach was based on the comparative evaluation of Closed-Set detectors (YOLOv10, YOLOv12, RT-DETR, RF-DETR) and OVD models (YOLO-World, OWL-ViT), using a fine-tuning and validation protocol on the public dataset "ThermoSolar PV".

The results of our experiments identified two cutting-edge architectures that define the state of the art for our specific task. First, the RF-DETR model established a new standard for accuracy (82.6% mAP@0.50), demonstrating the exceptional ability of its self-supervised backbone (DINOv2) to extract fine semantic features and better classify challenging defects. While its raw inference speed is not enough for real-time applications, its documented optimization potential via NVIDIA TensorRT leads it to attain the feasibility on embedded platforms. By integrating visual-linguistic interactions within a CNN architecture, YOLO-World achieves a very high level of accuracy (78.1% mAP@0.50) while maintaining an inference speed suitable for real-time applications.

References

Akram, M.W., Li, G., Jin, Y., Chen, X., Zhu, C., Zhao, X., Aleem, M., Ahmad, A., 2019. Improved outdoor thermography and processing of infrared images for defect detection in PV modules. *Solar Energy*, 190, 549-560. doi.org/10.1016/j.solener.2019.08.061

Al Mahdi, H., Leahy, P.G., Alghoul, M., Morrison, A.P., 2024. A Review of Photovoltaic Module Failure and Degradation Mechanisms: Causes and Detection Techniques. *Solar*, 4(1), 43-82. doi.org/10.3390/solar4010003

Barraz, Z., Sebari, I., Lamrini, N., Ait El Kadi, K., Ait Abdelmoula, I., 2025. A cascading decision system for enhanced anomaly classification of large-scale photovoltaic systems using Drone’s thermal data with class-imbalance problem. *Results in Engineering*, 25, 103876. doi.org/10.1016/j.rineng.2024.103876

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. *European Conference on Computer Vision (ECCV), 213-229*. doi.org/10.1007/978-3-030-58452-8_13

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y., 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection (Version 3). *arXiv*. doi.org/10.48550/ARXIV.2401.17270

Darabi, P., 2025. ThermoSolar-PV: A Curated Thermal Imagery Dataset for Anomaly Detection in Photovoltaic Modules. doi.org/10.13140/RG.2.2.12595.54564

De Oliveira, A.K.V., Aghaei, M., Rüther, R., 2020. Aerial infrared thermography for low-cost and fast fault detection in utility-scale PV power plants. *Solar Energy*, 211, 712-724. doi.org/10.1016/j.solener.2020.09.066

Hijjawi, U., Lakshminarayana, S., Xu, T., Piero Malfense Fierro, G., Rahman, M., 2023. A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations. *Solar Energy*, 266, 112186. doi.org/10.1016/j.solener.2023.112186

Li, B., Qi, W., Zhang, X., 2024. Research Advanced in Object Detection based on Deep Learning. *Highlights in Science, Engineering and Technology*, 119, 491-499. doi.org/10.54097/tax5ym24

Masita, K., Hasan, A., Shongwe, T., Hilal, H.A., 2025. Deep learning in defects detection of PV modules: A review. *Solar*

Energy Advances, 5, 100090. doi.org/10.1016/j.seja.2025.100090

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N., 2022. Simple Open-Vocabulary Object Detection with Vision Transformers (Version 2). *arXiv*. doi.org/10.48550/ARXIV.2205.06230

Sapkota, R., Cheppally, R.H., Sharda, A., Karkee, M., 2025. RF-DETR Object Detection vs YOLOv12: A Study of Transformer-based and CNN-based Architectures for Single-Class and Multi-Class Greenfruit Detection in Complex Orchard Environments Under Label Ambiguity (Version 1). *arXiv*. doi.org/10.48550/ARXIV.2504.13099

Tian, Y., Ye, Q., Doermann, D., 2025. YOLOv12: Attention-Centric Real-Time Object Detectors (Version 1). *arXiv*. doi.org/10.48550/ARXIV.2502.12524

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G., 2024. YOLOv10: Real-Time End-to-End Object Detection (Version 2). *arXiv*. doi.org/10.48550/ARXIV.2405.14458

Zefri, Y., Sebari, I., Hajji, H., Aniba, G., Aghaei, M., 2023. A layer-2 solution for inspecting large-scale photovoltaic arrays through aerial LWIR multiview photogrammetry and deep learning: A hybrid data-centric and model-centric approach. *Expert Systems with Applications*, 223, 119950. doi.org/10.1016/j.eswa.2023.119950

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J., 2023. DETRs Beat YOLOs on Real-time Object Detection (Version 3). *arXiv*. doi.org/10.48550/ARXIV.2304.08069