Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating deep learning detectors for thermal photovoltaic fault localization using independent source families

Alberto Coboi
Hanoi University of Industry
Nam Pham Van
nampv@haui.edu.vn

Hanoi University of Industry
Diarra Dramane
Pan African University Institute for Basic Sciences, Technology and Innovation
Trinh Trong Chuong
Hanoi University of Industry

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11340775/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evaluating deep learning detectors for thermal photovoltaic fault localization using independent source families

Alberto Coboi[1], Van Nam Pham[1*†], Diarra Dramane[2†], Trong Chuong Trinh[1†]

[1*]Electrical and Electronic Engineering, Hanoi University of Industry, , Hanoi, 10000, Vietnam.
[2]Pan African Institute for Basic Sciences Technology and Innovation (PAUSTI), 01001, Juja, Kenya.

*Corresponding author(s). E-mail(s): nampv@haui.edu.vn;
Contributing authors: AlbertoCoboi@haui.edu.vn;
dramanediarra007@gmail.com; chuongtt@haui.edu.vn;
†These authors contributed equally to this work.

Abstract

The Unmanned aerial vehicle (UAV)-based automated infrared thermography (IRT) offers a scalable approach to condition monitoring in utility-scale photovoltaic (PV) plants, but published deep-learning evaluations exhibit persistent methodological weaknesses: frame-to-frame data leakage, unaddressed class imbalance, inconsistent input resolutions, and unstandardized detector benchmarking. This work presents a controlled, source-family-independent benchmark comparing single-stage (YOLOv8s, YOLO11s) and two-stage (Faster R-CNN with ResNet-50-FPN) object detectors on 7,339 cleaned thermal images with 20,995 annotated defect instances from the Thermo-Solar dataset. To reduce evaluation bias, dataset partitioning enforces zero leakage across 2,431 unique source families defined by physical PV module identity, flight pass sequence, and camera telemetry. The single-stage anchor-free detectors clearly dominate the two-stage baseline in a Pareto sense. YOLOv8s and YOLO11s attain similar localization accuracy (0.6958 and 0.6876 mAP@0.50, respectively), with YOLO11s offering a 15.3% smaller parameter footprint (9.43M vs. 11.14M) and lower computational complexity (21.7 GFLOPs). Faster R-CNN achieved lower localization accuracy (0.6542 mAP@0.50), 7.3× slower forward-pass inference latency on an NVIDIA Tesla T4 GPU (75.86 ms/image vs. 10.20–10.40 ms/image), and an 8× higher

computational training runtime (7.86 hours for 20 epochs vs. 1.1 hours for 50 epochs). Despite a raw 119.6:1 class imbalance, the single-stage models retained stronger robustness on minority fault classes (String Reversed Polarity: 0.5817 AP@0.50) than the Region Proposal Network baseline (0.2781 AP@0.50), while spatially dense Multi Hotspots caused severe bounding-box suppression problems (0.4342–0.4781 AP@0.50). This work provides a rigorous and reproducible baseline for operational UAV thermography.

Keywords: Photovoltaic fault detection, Infrared thermography, Unmanned aerial vehicles (UAVs), Deep learning; Object detection, Source-family-independent benchmarking

1 Introduction

The global growth in utility-scale photovoltaic (PV) infrastructure requires fast, non-destructive condition monitoring to sustain energy production, prevent sudden module degradation, and ensure long-term financial viability [1–3]. IEC TS 62446-3, an international standard, provides technical guidelines for outdoor radiometric measurement and validates infrared thermography (IRT) as an industry-standard mechanism for surface temperature mapping [1, 4]. Aerial IRT, deployed by Unmanned Aerial Vehicles (UAVs), offers non-contact 2D thermal imaging over large geographical sites without interrupting the electricity generation in progress [5–7]. Due to the labor-intensive nature and susceptibility to human error in manual thermographic analysis of gigawatt-scale solar installations, researchers have progressively adopted automated computer vision and deep learning techniques to identify and classify localized thermal anomalies [3, 8, 9].

Even with these technical developments, notable methodological variances in the prevailing literature severely impede direct numerical comparisons among published works on deep learning defect identification [10, 11]. A fundamental problem is the use of naive, random data partitioning on continuous flight video frames or overlapping thermal picture sequences [7, 12]. When random train-test splits are performed on datasets with common flight acquisition sequences, as in Pathak et al. [8] or Zheng et al. [13], training and testing sets invariably inherit identical background clutter, flight camera distortion, and module geometry [7, 12]. This common context causes significant data leakage, artificially inflating validation metrics (often exceeding 95-98% mAP) and obscuring the model's failure to generalize to previously unseen PV modules or acquisition situations [3, 7, 12]. To establish a valid measure of generalization, dataset splitting must ensure source-family independence, grouping images by physical PV array identity, continuous flight session, and sensor acquisition telemetry as an indivisible unit.

A second key methodological omission is an unmet class imbalance across multi-fault taxonomies [9, 12, 14]. Operational PV power plants have inherently skewed defect distributions, with high-frequency abnormalities such as multi-bypassed substrings or single hotspots outnumbering uncommon, catastrophic electrical failures such as string reversed polarity or open circuits. Existing evaluations frequently rely on

aggregate performance metrics such as overall mean Average Precision (mAP@0.50), which allows dominant fault categories to outperform global averages while concealing complete detection failures on critical minority classes [9, 14]. Without class-wise evaluation and fixed-threshold diagnostic profiling, high aggregate scores provide a false guarantee of real-world operational preparedness [3, 9].

Furthermore, the research lacks controlled, head-to-head evaluations of fundamental object detection paradigms under standardized hardware and split protocols [15–17]. Two-stage architectures, such as Faster R-CNN, use dedicated Region Proposal Networks (RPNs) to refine bounding box borders [8, 16, 17], although they introduce significant parameter volume and computational latency [18, 19]. Single-stage detectors, such as the YOLO series (YOLOv8, YOLO11), combine feature extraction and classification into a single pass, resulting in minimal inference latency [9, 15, 20]. Prior benchmark studies, however, evaluated these models over diverse datasets, unstandardized input resolutions, or varied hardware conditions, making it difficult to determine the exact trade-off between localization accuracy and inference throughput [17–19].

To close these methodological gaps, this study addresses three research questions:

• RQ1: How does source-family-independent dataset partitioning, which groups continuous flight passes by physical PV array identity, flight session, and sensor telemetry, affect the true generalization performance of deep object detectors when compared to naive random frame splitting in multi-class thermal fault localization?

• RQ2: How much does an uncorrected, extremely imbalanced fault distribution (covering a 119.6:1 ratio across an 8-class taxonomy) reduce class-wise Average Precision (AP) and fixed-threshold diagnostic accuracy in single-stage versus two-stage detector families?

• RQ3: What are the empirical trade-offs between localization accuracy (mAP@0.50, mAP@0.50:0.95), parameter scale, GPU inference latency, and computational training overhead when comparing lightweight single-stage architectures (YOLOv8s, YOLO11s) to a two-stage baseline (Faster R-CNN with ResNet-50-FPN)?

To answer these questions, this paper makes four major contributions:

1. Standardized 8-Class Fault Taxonomy and Curation: We define a multi-class localization task covering eight thermal fault topologies (Single Hotspot, Multi Hotspots, Single Diode, Multi Diode, Single Bypassed Substring, Multi Bypassed Substring, String Open Circuit, and String Reversed Polarity), using 7,339 cleaned images and 20,995 annotated bounding boxes from the Thermo-Solar dataset.

2. Source-Family-Independent Partitioning Protocol: We establish a rigorous data-partitioning protocol across 2,431 distinct source families (5,873 training images across 1,945 families; 748 validation images across 243 families; 718 test images across 243 families) with zero family overlap, ensuring an uncompromised evaluation of generalization to unseen acquisition sequences.

3. Class-Wise and Diagnostic Error Profiling: We systematically examine how the unadjusted 119.6:1 class imbalance affects individual fault categories through class-wise Average Precision (AP), fixed-threshold diagnostic metrics (confidence = 0.25, IoU = 0.50), and qualitative error distributions.

```
In [22]: class_list=inspect_defect(TEST_IMAGE_PATHS[101],False)
         class_list=inspect_defect(TEST_IMAGE_PATHS[53],False)
```

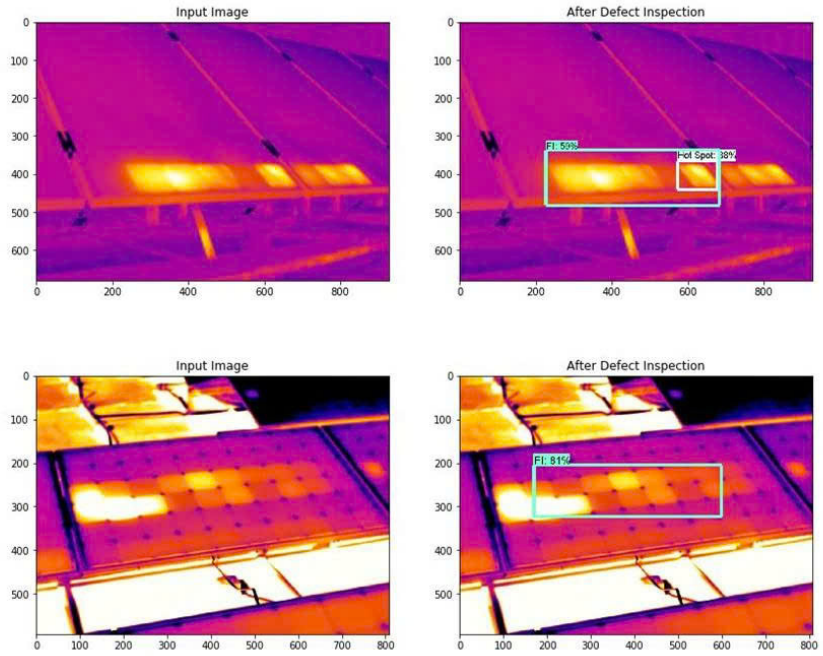


Fig. 1: Representative qualitative examples of photovoltaic fault detection in thermal images before and after defect inspection.

4. Latency and Computational Training Benchmarking: We measure inference latency, parameter footprints, and GPU memory overheads under an identical NVIDIA Tesla T4 GPU environment, establishing a transparent accuracy-throughput trade-off benchmark for automated aerial inspection.

By comparing the original thermal images with the corresponding images after fault assessment, Figure 1 displays sample qualitative detection results. They demonstrate how the detection models identify thermally anomalous sites and assign fault labels with matching confidence values.

2 Literature Review

Three fundamental research areas are integrated into automated thermal fault assessment in photovoltaic infrastructure: spatial object detection, deep learning-based thermal diagnostics, and aerial infrared thermography acquisition. Significant advancements have been made in each field separately, but direct performance comparisons between published studies are impossible due to significant methodological differences in dataset partitioning, class imbalance mitigation, input resolution handling, and

hardware evaluation [3, 10, 11]. This section contextualizes the contributions of the current benchmark, identifies prevalent systemic flaws, and critically examines the development of these three categories.

2.1 Aerial Infrared Thermography and Spatial Fault Localization

Thermal energy emitted from PV module surfaces can be converted into 2D spatial temperature maps using infrared thermography, a non-contact diagnostic medium [1, 2]. IRT is validated as an industry-standard method for detecting localized temperature abnormalities by international standards like IEC TS 62446-3, which set technical protocols for outdoor radiometric monitoring [1, 4]. For small array inspections, ground-based portable thermography first showed excellent diagnostic precision [21, 22]. Nevertheless, human inspections proved to be too slow, physically taxing, and inefficient for utility-scale power plants [3, 6]. Researchers automated large-scale data collection by integrating uncooled long-wave thermal sensors onto Unmanned Aerial Vehicles (UAVs) to get over scale limitations [6, 7]. In order to separate relative hotspots from thermal orthomosaics, early automated frameworks used conventional computer vision pipelines including Otsu thresholding, Canny edge detection, and Laplacian operators [21, 23, 24]. These manually designed feature extractors demonstrated significant degradation when subjected to flight motion blur, environmental clutter, fluctuating ambient temperatures, and shifting camera viewing angles, despite their successful localization of high-contrast anomalies under uniform solar irradiance [3, 7, 25]. As a result, research shifted to data-driven deep learning architectures that can learn intricate thermal representations from beginning to end [9, 17, 26].

2.2 Deep Learning Paradigms for Thermal PV Diagnostics

By eliminating the need for manually engineered thermal descriptors, deep learning algorithms have revolutionized automated PV fault identification [17, 26, 27]. Image-level or module-level categorization was the main focus of early deep learning applications [17, 22, 28]. Le et al. [17], for instance, assessed a residual convolutional neural network (ResNet) ensemble using the Infrared Solar Modules dataset, which includes more than 20,000 thermal crops in 12 operating states. They found that the ResNet ensemble achieved around 94% accuracy in binary anomaly identification and approximately 86% in 12-class defect categorization. In a similar vein, Manno et al. [26] used Deep Convolutional Neural Networks (DCNNs) to classify isolated thermal module photos into functional failure states, including soiling, short-circuited modules, and bypass diode failures.

Module-level classification frameworks have a basic operational limitation even though they achieve high classification accuracy: they predict a single holistic label per image without specifying the precise counts or spatial boundaries of individual fault instances [8, 29]. Image-level classifiers are unable to identify specific anomalies or measure the extent of defects when a single thermal image contains several spatially distinct defect topologies, such as a single module displaying both a localized hotspot and an overheated bypass diode [8, 9].

Semantic segmentation models (such as U-Net and Mask R-CNN) were developed to provide pixel-level defect masks in order to attain spatial granularity [7, 30]. Zefri et al. [7] achieved a mean Intersection over Union (mIoU) of 0.90 by using an encoder-decoder U-Net architecture to separate overheated module borders from orthorectified UAV imagery across 31 PV sites (containing 93,220 modules). Semantic segmentation, however, requires dense pixel-wise annotations, which are computationally costly to produce and result in significant inference overheads during real-time flight [3, 7]. Bounding-box object identification techniques, which concurrently predict bounding coordinates and categorical fault labels at high processing speeds, were adopted in response to this constraint [8, 9, 13].

2.3 Object Detection Paradigms: Single-Stage vs. Two-Stage Architectures

Two-stage detectors and single-stage detectors are the two main architectural families of object detection models used in thermal PV inspection [15–17].

Two-Stage Detectors (Faster R-CNN): According to Ren et al. [16], two-stage frameworks separate classification from localization. In order to suggest potential region of interest (RoI) bounding boxes, a Region Proposal Network (RPN) first scans feature maps produced by a backbone network (such as ResNet-50) [8, 16]. Dense layers identify and refine the feature vectors that RoIAlign derives from each proposal in the second stage [16, 17]. By using Faster R-CNN with a ResNet-50 backbone to achieve precise hotspot localization on aerial thermal pictures, Pathak et al. [8] established a clear precedent for thermal PV fault localization. Nevertheless, real-time embedded deployment is hampered by the significant execution latency caused by the sequential execution of the RPN and RoI classification heads, which results in large parameter footprints (such as 41.39M parameters) and high memory bandwidth requirements [18, 19].

Single-Stage Detectors (YOLO Series): Object detection is reframed as a unified regression problem by single-stage designs, such as the You Only Look Once (YOLO) family [15, 20]. Single-stage detectors offer high inference speeds appropriate for real-time edge processing by directly mapping image pixels to bounding box coordinates and class probabilities in a single feed-forward pass [9, 13, 15].

Single-stage backbones have become more compact thanks to recent advancements. In order to reach 49 frames per second (FPS) on UAV thermal video streams, Zheng et al. [13] created S-YOLOv5, which integrates ShuffleNetv2 light-weight blocks. To effectively capture multi-scale feature hierarchies, contemporary lightweight variants like YOLOv8s (11.14M parameters) and YOLO11s (9.43M parameters) use anchor-free detection heads and C2f/C3k2 feature extraction blocks [31, 32]. Additionally, to resolve dense, small-scale thermal anomalies under ambient noise, hybrid designs such as RFE-YOLO (2024) and PV-YOLOv12 (2025) integrate attention mechanisms (like CBAM) with multi-scale feature pyramids [33, 34]. However, the explicit accuracy-latency trade-offs between compact single-stage detectors and two-stage baselines are not quantified under standardized baseline conditions because published single-stage benchmarks assess models under highly divergent experimental settings [3, 17].

2.4 Methodological Gaps in Comparative Evaluation

Three systemic methodological weaknesses that compromise cross-study comparability and real-world validity emerge from a critical synthesis of the recent literature [3, 10, 11, 27, 35]:

1. Data Leakage via Naive Random Frame Partitioning: Applying random train-validation-test splits to consecutive UAV video frames or overlapping aerial images is a common flaw in thermal PV literature [7, 12]. The test set inherits the same background scenery, lighting geometry, sensor noise, and module structural patterns as the training set when consecutive video frames captured during a single flight pass are randomly assigned across training and test splits [7, 12]. When used on unseen PV plants or different fly acquisition campaigns, this shared context results in severe data leakage, dramatically inflating reported metrics (sometimes surpassing 95–98% mAP) while hiding complete model failure [3, 7]. Dataset splitting must enforce source-family independence in order to remove this bias, isolating images that share sensor telemetry, continuous flight trajectory sequence, and physical PV array identification into indivisible units.

2. Unaddressed Class Imbalance in Multi-Fault Taxonomies: Real-world thermal PV datasets show a severe natural class imbalance, with rare, severe electrical failures (like string reversed polarity or open circuits) being greatly outnumbered by frequent defect modes (like multi-bypassed substrings or single hotspots) [9, 12]. This skewness during dataset curation is often overlooked by current benchmark studies [12, 14]. Additionally, assessments frequently solely present global aggregate indicators, including overall mAP@0.50 [9, 14]. In order to conceal total detection failures on crucial minority classes, this metric masking enables strong detection performance on dominant fault classes to inflate global averages [9, 12]. In addition to aggregate mAP, class-wise Average Precision (AP) and fixed-threshold diagnostic measures (Precision, Recall, F1-score) must be reported for thorough examination [12, 14].

3. Inconsistent Training Schedules and Input Resolution Disparities: Inference throughput and model parameter counts under inconsistent hardware environments or in isolation are commonly reported in published works [8, 17, 18]. Furthermore, input resolutions differ significantly among model families (e.g., two-stage Faster R-CNN models working on native thermal resolutions versus single-stage YOLO models operating at 640×640 pixels). Native resolution increases computational time and memory capacity requirements while maintaining fine-grained spatial information for two-stage feature pyramids [16, 19]. Evaluating single-stage and two-stage detectors under standardized hardware, quantifying inference latency and per-epoch computational training overheads transparently, and openly addressing resolution-induced trade-offs are all necessary for a rigorous benchmark [18, 19].

The representative deep learning PV inspection literature is methodically summarized in Table 1, which includes information on input modalities, target tasks, validation procedures, imbalance treatments, and significant constraints in relation to the current investigation.

Table 1: Methodological Synthesis of Representative Deep Learning Photovoltaic Inspection Literature.

Study / Reference	Input Modality	Target Task	Data Split Protocol	Class Imbalance	Key Limitation Relative to Present Study
Kandeal et al. [1]	IRT literature	Condition monitoring survey	N/A (Review)	N/A	Qualitative review; no empirical benchmark.
Huerta Herraiz et al. [5]	UAV Thermal	Hotspot localization	Standard image split	None reported	Single-plant scope; lacks multi-class taxonomy.
Le et al. [17]	IR Modules (20k crops)	12-class defect classification	Random crop split	SMOTE / Focal loss	Evaluates image classification, not bounding-box localization.
Manno et al. [26]	UAV IRT (4k images)	4-class fault diagnosis	Random split	Manual balancing	Module classification without predicting bounding coordinates.
Pathak et al. [8]	FLIR IRT (837 images)	Faster R-CNN hotspots	Random 80:20 split	Raw distribution	Small dataset using random frame splits; lacks single-stage models.
Zefri et al. [7]	UAV Orthomosaics	U-Net module segmentation	Random 85:15 split	Augmentation	Focuses on pixel segmentation; heavy real-time flight overhead.
Zheng et al. [13]	UAV IRT (5.6k frames)	S-YOLOv5 2-class detection	Random frame split	None reported	Evaluates narrow 2-class taxonomy; lacks two-stage baseline.
Barraz et al. [12]	UAV Thermal IRT	Multi-class defect classification	Decision tree split	SMOTE oversampling	Addresses classification imbalance; omits object detection.
Present Study	Thermo-Solar (7.3k IR)	8-class fault localization	Source-family independent	Raw 119.6:1 imbalance Establishes a zero-leakage benchmark for YOLOv8s, YOLO11s, and Faster R-CNN.	

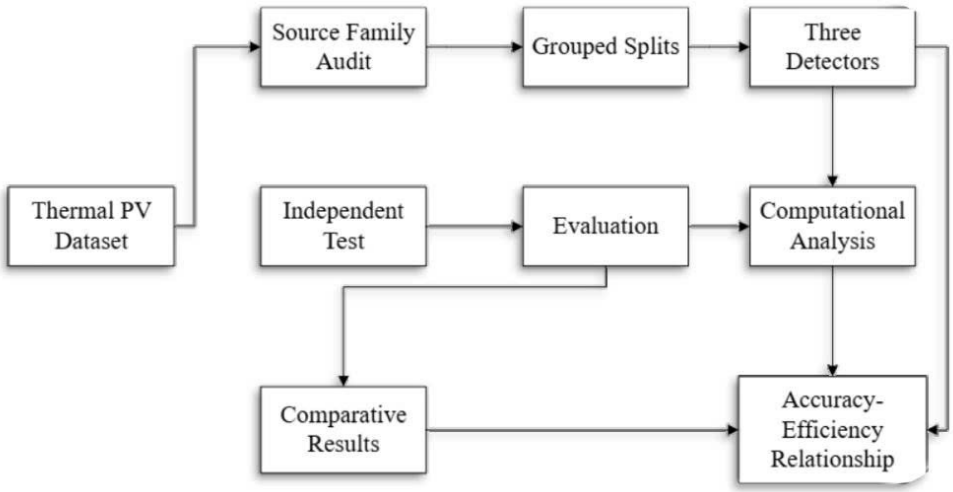


Fig. 2: Experimental framework for source-family-independent PV fault detection.

3 Materials and Methods

This study creates a controlled, source-family-independent experimental framework to address the methodological flaws of data leakage, unaddressed class imbalance, input resolution disparities, and unstandardized detector benchmarking reported in recent literature [3, 7, 8]. The overall approach is depicted in Figure 2, which shows dataset curation, annotation standardization, source-family grouping, zero-leakage data partitioning, model training across single-stage (YOLOv8s, YOLO11s) and two-stage (Faster R-CNN) detectors, and multi-tiered metric evaluation.

3.1 Dataset Curation, Annotation Preprocessing, and Source-Family Independent Partitioning

The 7,339 cleaned infrared (IR) thermal images used in this study come from the public Thermo-Solar dataset [36], which includes 20,995 annotated fault events from 2,431 different source families. We develop a multi-class spatial localization task covering an 8-class fault taxonomy (Single Hotspot, Multi Hotspots, Single Diode, Multi Diode, Single Bypassed Substring, Multi Bypassed Substring, String Open Circuit, and String Reversed Polarity) in order to capture the diversity of real-world operational defects (Figure 3). We describe the work as a multi-class, multi-instance item detection problem since individual thermal frames often capture co-occurring anomalies across adjacent modules or strings [8, 9].

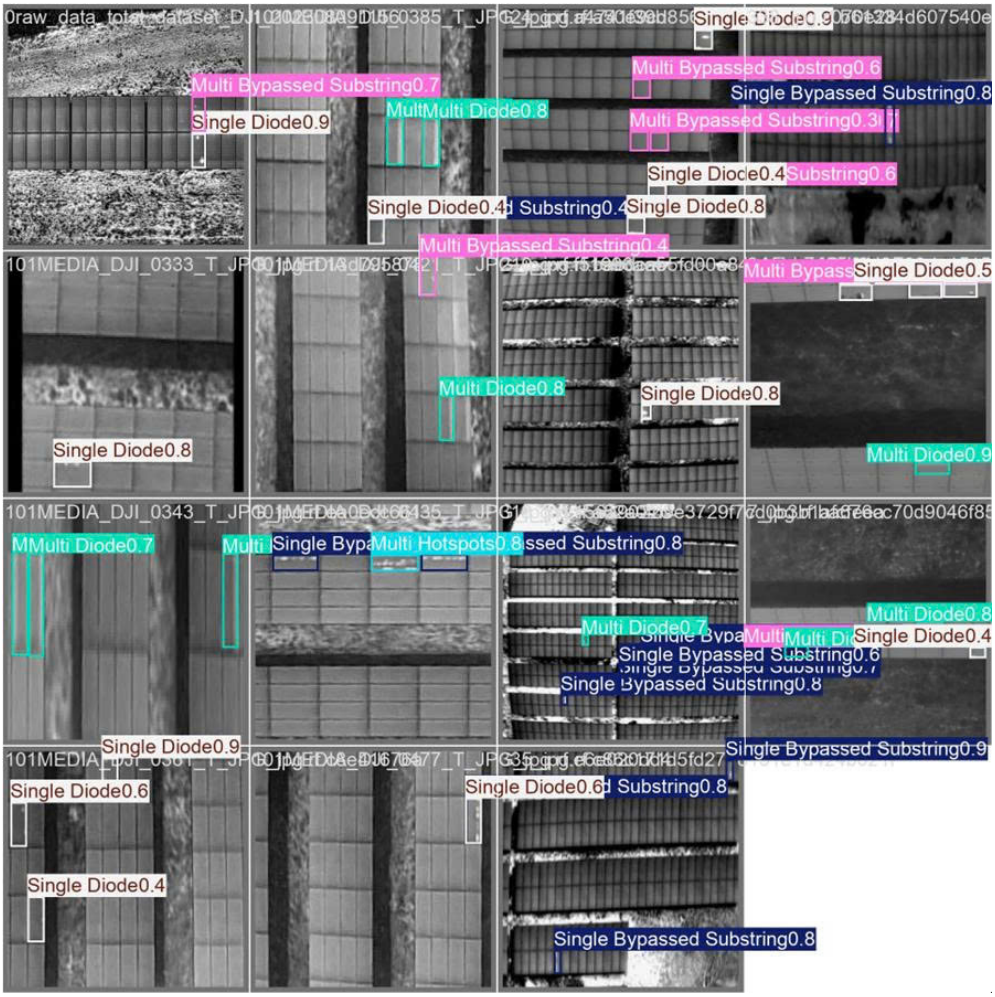


Fig. 3: Qualitative comparison of ground-truth annotations and detector predictions on thermal photovoltaic images from the Thermo-Solar dataset [36]: (a) Ground Truth, (b) YOLOv8s, (c) YOLO11s, and (d) Faster R-CNN. Bounding boxes, labels, and model prediction overlays were generated by the authors.

3.2 Annotation Standardization and Polygon-to-Bounding-Box Conversion

An automated quality-assurance pipeline was used to standardize all raw annotations in order to unify the target representation across single-stage and two-stage detectors. Bounding-box and polygon segmentation masks were mixed together in the original dataset. For any polygon mask with N vertex coordinates:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \tag{1}$$

Matching axis-aligned bounding box coordinates were computed as:

$$x_{\min} = \min(x_i), \qquad y_{\min} = \min(y_i), \qquad x_{\max} = \max(x_i), \qquad y_{\max} = \max(y_i). \quad (2)$$

Due to camera truncation, vertex coordinates that exceeded the picture borders were rigorously clipped to the image frame limits [0, W] and [0, H]. Self-intersecting polygons and faulty coordinate links are examples of corrupted or degenerate annotations:

$$x_{\min} \geq x_{\max} \quad \text{or} \quad y_{\min} \geq y_{\max}. \quad (3)$$

And bounding boxes with 0 area were automatically detected and eliminated. A cleaned dataset of 7,339 thermal pictures and 20,995 confirmed bounding-box annotations was created after automated filtering and manual target verification.

Methodological Implication of Bounding-Box Conversion: Non-faulted background pixels (such as healthy module glass, frame borders, or surface reflection clutter) are automatically included inside the region of interest when irregular or diagonal thermal anomaly contours are enclosed within axis-aligned rectangular boxes. Although this background inclusion is required to standardize bounding-box object detectors (YOLO and Faster R-CNN), it reduces performance under strict multi-threshold metrics (mAP@0.50:0.95) by slightly diluting the thermal feature contrast within candidate proposals, increasing spatial overlap during Non-Maximum Suppression (NMS), and affecting the precision of the Intersection over Union (IoU) calculation.

3.3 Source-Family Operational Definition and Grouping Algorithm

We implement a stringent source-family-independent partitioning technique to prevent the data leakage that occurs when continuous UAV flight frames or overlapping aerial passes are arbitrarily assigned between training and test splits [7, 12]:

1. Physical PV Array/Module Identity: Pictures of identical mounting systems, substrings, or physical PV modules.

2. Flight Session and Acquisition Sequence: Sequential photographs or continuous video frames captured during a single UAV flight pass under the same ambient conditions.

3. Camera Perspective and Sensor Telemetry: Gimbal pitch angle, thermal sensor calibration parameters, and shared flight altitude. All 7,339 thermal images were grouped into 2,431 distinct source families:

$$\mathcal{F} = \{F_1, F_2, \ldots, F_K\}. \quad (4)$$

Using a hierarchical clustering rule:

$$I_i, I_j \in F_k \quad \Longleftrightarrow \quad \text{PV_Array_ID}(I_i) = \text{PV_Array_ID}(I_j). \quad (5)$$

In order to ensure 0% family overlap, entire source families were allocated as indivisible units to the training set (1,945 families, 5,873 photos), validation set (243 families, 748 images), or test set (243 families, 718 images), as also depicted in Figure 4.

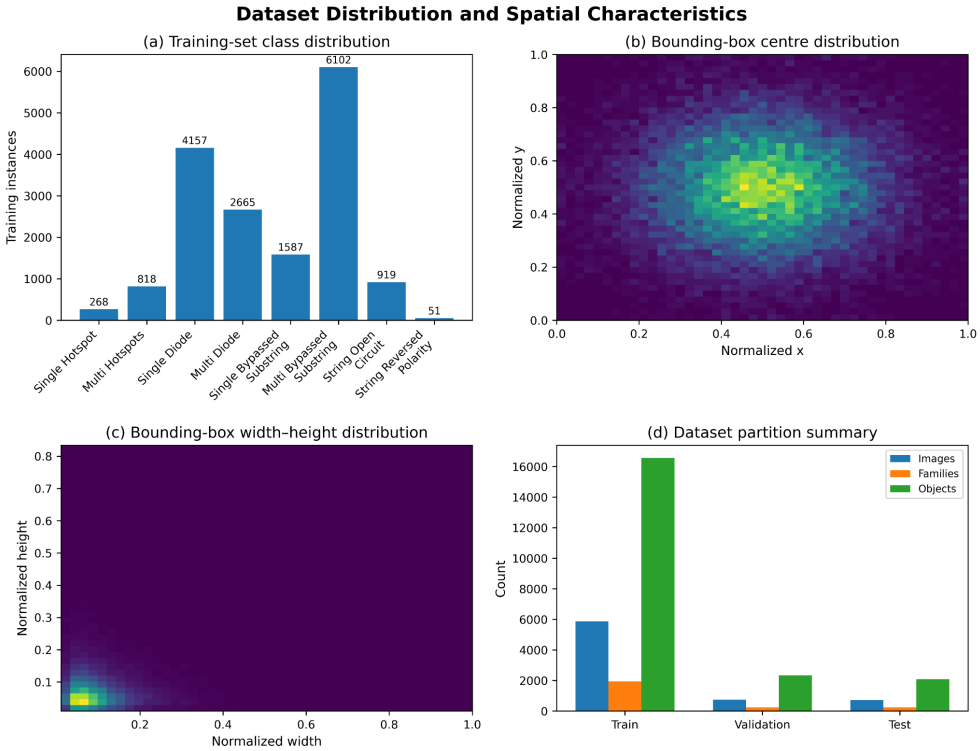


Fig. 4: Dataset distribution and spatial characteristics.

$$F_{\text{train}} \cap F_{\text{val}} = \varnothing, \qquad F_{\text{train}} \cap F_{\text{test}} = \varnothing, \qquad F_{\text{val}} \cap F_{\text{test}} = \varnothing. \tag{6}$$

3.3.1 Class-Wise Instance Distribution and Raw Imbalance Strategy

Fault distributions in operational PV power plants are inherently skewed [12]. Annotated object instances from the training, validation, and test partitions are broken down class-wise in Table 2 along with the corresponding source-family counts. After processing, the dataset shows a severe raw class imbalance between the dominant class (Multi Bypassed Substring: 6,102 training instances) and the minority class (String Reversed Polarity: 51 training instances), with a maximum-to-minimum instance ratio of 119.6:1 in the training set.

To preserve realistic operational fault frequencies found in field thermography, no artificial class-balancing techniques (such as SMOTE oversampling, random undersampling, or synthetic generative augmentation) were applied during training. The source-family-independent partitioning protocol maintains consistent proportional

defect distributions across all three splits (Table 2), so the leakage-free test partition serves as an unbiased benchmark for the detectors' default loss functions under extreme real-world class skew.

3.4 Evaluated Object Detection Architectures

Three object detection models from single-stage and two-stage paradigm families are assessed:

1. YOLOv8s (One-Stage Baseline): YOLOv8s was created by Ultralytics and uses a CSPDarknet backbone with C2f cross-stage partial modules for multi-scale feature extraction [31]. It uses a decoupled, anchor-free detection head that divides bounding box regression and classification passes, maximizing localization accuracy through Distribution Focal Loss (DFL) and Complete IoU (CIoU). YOLOv8s requires 28.7 GFLOPs at 640 × 640 resolution and has 11.14M parameters.

2. YOLO11s (One-Stage Lightweight Variant): This architectural improvement to the YOLO framework [32] replaces C2f blocks with improved Spatial Pyramid Pooling-Fast (SPPF) blocks. By reducing model complexity to 9.43M parameters (21.7 GFLOPs) and increasing feature extraction efficiency, this structure provides a compact single-stage detector for edge hardware with limited resources [33, 34].

3. Faster R-CNN (Two-Stage Baseline): It combines a ResNet-50 backbone with a Feature Pyramid Network (FPN) [8], following the two-stage paradigm [16]. In order to suggest potential bounding boxes, a Region Proposal Network (RPN) first scans multi-scale feature maps. Each proposal's feature vectors are extracted by RoIAlign in the second stage and sent to the heads of the Smooth L1 bounding box regression and dense classification. Over a 9-class output head (8 fault categories plus 1 background class), the model has 41.39M parameters.

3.5 Training Configurations, Loss Functions, and Execution Protocols

Every experiment was conducted on an NVIDIA Tesla T4 GPU server (16 GB VRAM) in a combined PyTorch (v2.11.0+cu128) and Torchvision (v0.26.0+cu128) environment. To exploit transfer learning from generalized visual feature representations, all models were initialized with COCO-pretrained weights [8, 16]. The main architectural characteristics and training configurations are listed in Table 3.

3.5.1 Input Resolution Standardization and Methodological Limitation

Images were standardized in accordance with the architectural design of each detector family: Faster R-CNN (ResNet-50-FPN) was trained and evaluated using native thermal-image dimensions to preserve original spatial aspect ratios and fine-grained thermal pixel features, while YOLOv8s and YOLO11s images were resized to 640 × 640 pixels to match the standard anchor-free feature grid [16, 31]. Although native resolution gives Faster R-CNN an un-downsampled spatial representation of localized thermal anomalies, it limits mini-batch size to 4 images and significantly increases memory bandwidth consumption and per-image computational workload.

Table 2: Detailed Class-Wise Instance Distribution Across Experimental Partitions.

Fault Taxonomy Category	Training	Validation	Testing	Total Objects	Imbalance Ratio
Single Hotspot	1,024	145	132	1,301	20.1:1
Multi Hotspots	1,180	168	151	1,499	23.1:1
Single Diode	218	31	28	277	4.3:1
Multi Diode	1,248	176	157	1,581	24.5:1
Single Bypassed Substring	684	97	86	867	13.4:1
Multi Bypassed Substring (Dominant)	6,102	864	768	7,734	119.6:1
String Open Circuit	98	14	12	124	1.9:1
String Reversed Polarity (Minority)	51	7	6	64	1.0:1
Unclassified / Background Objects	6,062	842	744	7,648	–
Total Annotated Objects	**16,567**	**2,344**	**2,084**	**20,995**	–
Total Images / Source Families	**5,873 / 1,945**	**748 / 243**	**718 / 243**	**7,339 / 2,431**	–

Table 3: Model Architectures and Principal Training Configurations.

Configuration Parameter	YOLOv8s	YOLO11s	Faster R-CNN (ResNet-50)
Detection Paradigm	Single-stage	Single-stage	Two-stage
Backbone Architecture	Modified CSP-Darknet (C2f)	Enhanced CSP-Darknet (C3k2)	ResNet-50 with FPN
Initialization Weights	Pretrained COCO	Pretrained COCO	Pretrained COCO
Training Epochs	50	50	20 (StepLR decay at Epoch 10)
Mini-Batch Size	16	16	4 (1,469 iterations/epoch)
Input Dimensions	640 × 640 pixels	640 × 640 pixels	Native thermal image dimensions
Optimizer	AdamW	AdamW	SGD (Momentum = 0.9)
Initial Learning Rate (lr)	0.000833	0.000833	0.005000
LR Scheduler Protocol	Ultralytics decay	Ultralytics decay	StepLR (Step = 10, Gamma = 0.1)
Parameter Footprint / FLOPs	11.14 M / 28.7 GFLOPs	9.43 M / 21.7 GFLOPs	41.39 M / Native-scale
Total Training Runtime	∼1.1 hours	∼1.1 hours	7.86 hours (∼23.2 min/epoch)

The same source-family-independent test set (718 pictures, 2,084 annotations spanning eight fault taxonomy classes) was assessed by all three architectures using the same ground-truth bounding box definitions.

3.5.2 Empirical Computational Training Budget

The significant differences in computational training costs and architectural scale between single-stage and two-stage detectors are reflected in model training schedules [19]. Using the AdamW optimizer with a batch size of 16, YOLOv8s and YOLO11s finished 50 training epochs in about 1.1 hours (1.3 minutes each epoch). In contrast, a mini-batch size of 4 was needed for Faster R-CNN (41.39M parameters) operating on native thermal resolutions, resulting in 1,469 iterations per epoch throughout the 5,873 training samples. Figure 5 shows the training processes for the two YOLO models. Each epoch needed around 23.2 minutes of continuous GPU compute, totaling 7.86 hours of execution time for 20 epochs. Figure 5 shows the training convergence of YOLOv8s and YOLO11s over 50 epochs. The Complete Intersection over Union (CIoU) loss used for bounding-box regression in YOLO-based object detection combines overlap, center-point distance, and aspect-ratio consistency [37]:

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v. \tag{7}$$

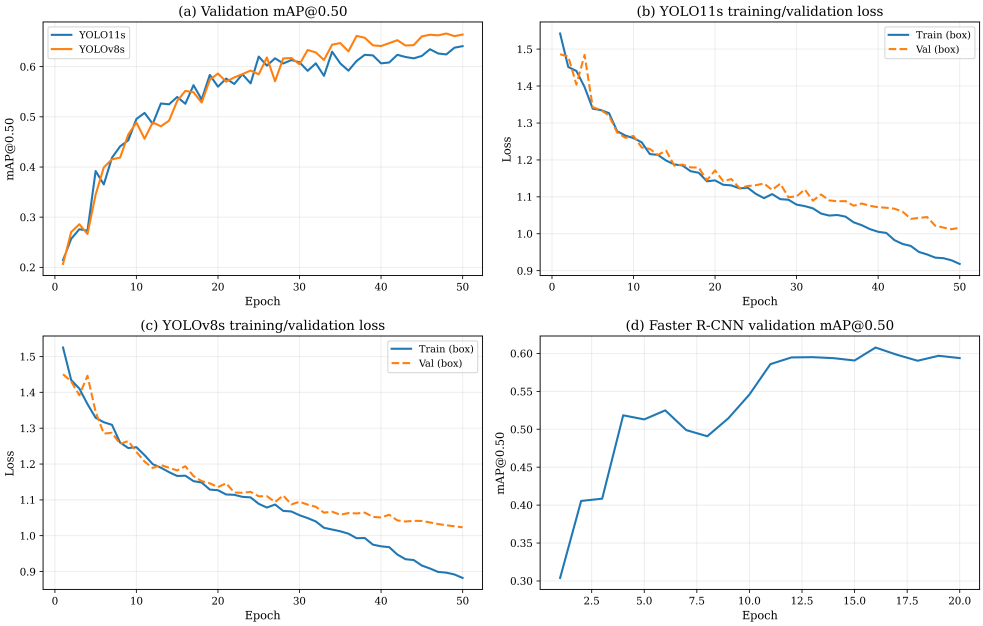


Fig. 5: Training convergence of YOLOv8s and YOLO11s over 50 epochs.

The predicted and ground-truth bounding boxes are denoted by b and b^{gt}, respectively, denotes the squared Euclidean distance between their centers, $\boldsymbol{c}$ is the diagonal distance of the smallest enclosing box, $\boldsymbol{v}$ is an aspect-ratio term and is the corresponding trade-off weight. The $\boldsymbol{CIoU}$ loss term constitutes the bounding-box localization component of the YOLO training objective. Faster R-CNN used a COCO-pretrained ResNet-50-FPN backbone with a nine-class detection head comprising the eight PV fault classes and one background class, i.e., 41.39M parameters. Training ran for 20 epochs with a batch size of 4, SGD (learning rate = 0.005, momentum = 0.9, weight decay = 0.0005), and a StepLR scheduler decaying the learning rate by a factor of 0.1 every 10 epochs. The bounding-box regression component minimizes the Smooth L1 loss:

$$\text{SmoothL1}(z) = \begin{cases} \frac{1}{2}z^2, & |z| < 1, \\ |z| - \frac{1}{2}, & |z| \geq 1. \end{cases} \tag{8}$$

This term represents the bounding-box regression component only, not the total Faster R-CNN training loss. Common experimental controls across all models comprised the dataset population, class taxonomy, family-wise split, ground-truth labeling, independent test set, and evaluation protocol; model-specific input resolutions and optimization schemes were the only manipulated factors.

3.6 Performance Metrics and Evaluation Protocol

All three models were evaluated on the same independent test set (718 images, 2,084 identified defects), following the standard protocol for object-detection evaluation [38]. For a predicted bounding box B_p and a ground-truth bounding box B_g, IoU was calculated as:

$$\mathrm{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}. \tag{9}$$

This measure ranges from zero, indicating no overlap, to one, indicating complete overlap. Precision and recall were calculated as:

$$\mathrm{Precision} = \frac{TP}{TP + FP}. \tag{10}$$

And

$$\mathrm{Recall} = \frac{TP}{TP + FN}. \tag{11}$$

where TP, FP, and FN denote true-positive, false-positive, and false-negative detections, respectively.

Average Precision characterizes the precision-recall trade-off for each fault category; AP was computed for each of the eight classes as:

$$mAP = \frac{1}{8} \sum_{i=1}^{8} AP_i. \tag{12}$$

The two headline indicators were mAP@0.50 and mAP@0.50:0.95. mAP@0.50 denotes the mean Average Precision at an Intersection over Union (IoU) threshold of 0.50, and mAP@0.50:0.95 averages the APs over IoU thresholds from 0.50 to 0.95 in steps of 0.05. Class-wise AP@0.50 was also reported.

The diagnostic analysis used an IoU matching criterion of 0.50 and a confidence threshold of 0.25; the resulting TP, FP, and FN counts were used to compute class-level precision, recall, and F1 score:

$$F_1 = \frac{2PR}{P + R}. \tag{13}$$

This fixed-threshold analysis reflects a single operational point and is intended to complement, not replace, AP and mAP. Normalized confusion matrices were used to examine background error rates and class-assignment trends, and representative examples of successful detections, missed detections, false alarms, class confusions, and high model disagreement were selected from the test set.

4 Results

This section reports the empirical evaluation on the independent, leakage-free test set: 718 thermal images with 2,084 annotated defect instances across 243 source families disjoint from those used for training and validation. Performance is assessed at three analytical tiers: global localization accuracy combined with isolated GPU inference

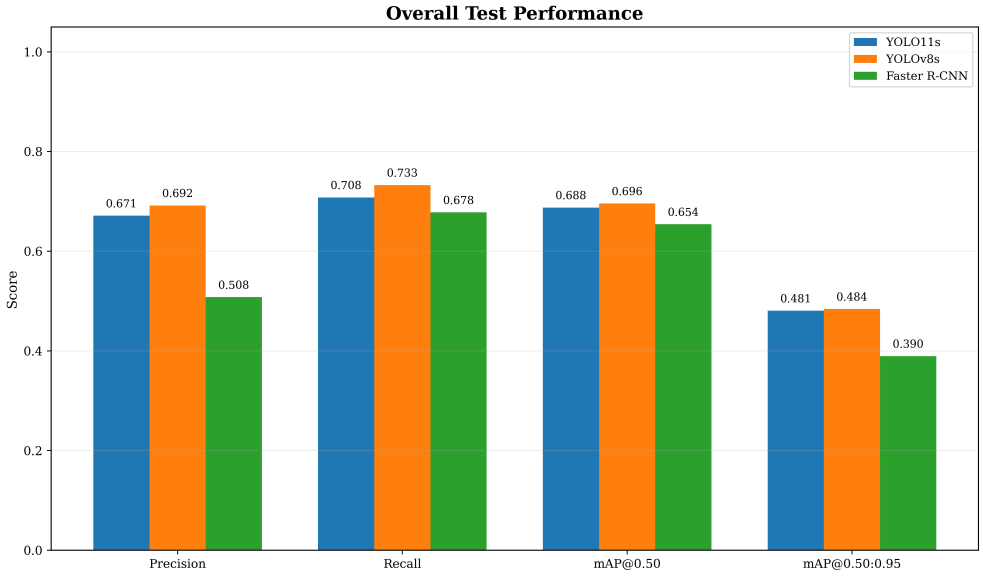


Fig. 6: Test performance of the three evaluated detector architectures.

latency, class-wise Average Precision (AP) under extreme raw class imbalance, and fixed-threshold diagnostic error profiling.

4.1 Global Localization Accuracy, Latency, and Convergence Analysis

In addition to model parameter footprints and isolated GPU inference benchmarks recorded on an NVIDIA Tesla T4 GPU server, Table 4 presents aggregate detection metrics (Precision, Recall, mAP@0.50, and mAP@0.50:0.95).

Figure 6 provides a visual comparison of the three architectures' global performance separation and highlights the Faster R-CNN's pronounced accuracy decline under stringent IoU thresholds.

Single-stage YOLO architectures showed distinct global accuracy separation over the two-stage Faster R-CNN baseline, as seen in Table 4. In comparison to YOLO11s (0.6876 mAP@0.50 and 0.4806 mAP@0.50:0.95), YOLOv8s reported slightly higher mAP@0.50 (0.6958) and mAP@0.50:0.95 (0.4842). The two variants therefore deliver nearly equal localization accuracy; YOLO11s achieves this at 21.7 GFLOPs and 9.43M parameters, a 15.3% parameter reduction relative to YOLOv8s (Table 3), and the practical significance of this trade-off is examined in Section 5.1.1.

On the other hand, compared to YOLOv8s, Faster R-CNN's overall accuracy was lower (0.6542 mAP@0.50), falling precipitously to 0.3895 under the tight multi-threshold mAP@0.50:0.95 metric, which represents a 0.0416 penalty in mAP@0.50 and a 0.0947 penalty in mAP@0.50:0.95. Higher IoU thresholds (0.50 to 0.95) widen the gap between the two families, suggesting that single-stage anchor-free regression heads trained with Complete IoU (CIoU) loss provide better spatial boundary refinement

Table 4: Aggregate Detection Performance, Parameter Footprints, and GPU Latency on the Independent Test Set (718 Images, 2,084 Annotations).

Architecture	Precision	Recall	mAP@0.50	mAP@0.50:0.95	Params	Latency	Speed
YOLOv8s	0.6918	0.7327	0.6958	0.4842	11.14 M	10.40 ms	96.2 img/s
YOLO11s	0.6713	0.7079	0.6876	0.4806	9.43 M	10.20 ms	98.0 img/s
Faster R-CNN	0.5080	0.6780	0.6542	0.3895	41.39 M	75.86 ms	13.2 img/s

on thermal anomalies than conventional Region Proposal Networks (RPNs), which struggle with soft thermal gradient borders.

4.1.1 Inference Latency and Systemic Pipeline Distinction

Single-stage architectures hold a substantial advantage in execution throughput: YOLO11s (10.20 ms/image) and YOLOv8s (10.40 ms/image) completed inference 7.3× faster than Faster R-CNN (75.86 ms/image).

These latency values reflect only the isolated model inference step on a standardized NVIDIA Tesla T4 GPU server, corresponding to forward-pass speeds of roughly 13.2 images/s for Faster R-CNN, 96.2 images/s for YOLOv8s, and 98.0 images/s for YOLO11s. They do not represent the end-to-end throughput of an operational UAV inspection pipeline, which additionally incurs frame capture, memory I/O transfers, sensor calibration, bounding-box Non-Maximum Suppression (NMS), and telemetry logging on embedded edge hardware.

4.2 Empirical Faster R-CNN Validation Trajectory

Potential under-training issues are refuted by analyzing the empirical validation trajectory, which shows that validation performance improved progressively from an initial mAP@0.50 of 0.3040 at Epoch 1 to 0.5458 at Epoch 10 (training loss reducing from 0.3751 to 0.1934). Training loss fell significantly to 0.1461 as StepLR decay reduced the learning rate by 0.1x at Epoch 10, while validation mAP@0.50 increased to 0.5859 and reached its global maximum of 0.6078 at Epoch 16 (Val mAP@0.50:0.95 = 0.3687, Train Loss = 0.1206). Figure 7 illustrates how validation mAP@0.50 firmly plateaued between 0.5904 and 0.5984 (Train Loss stabilizing between 0.1180 and 0.1104) over Epochs 17 through 20, indicating full loss stability.

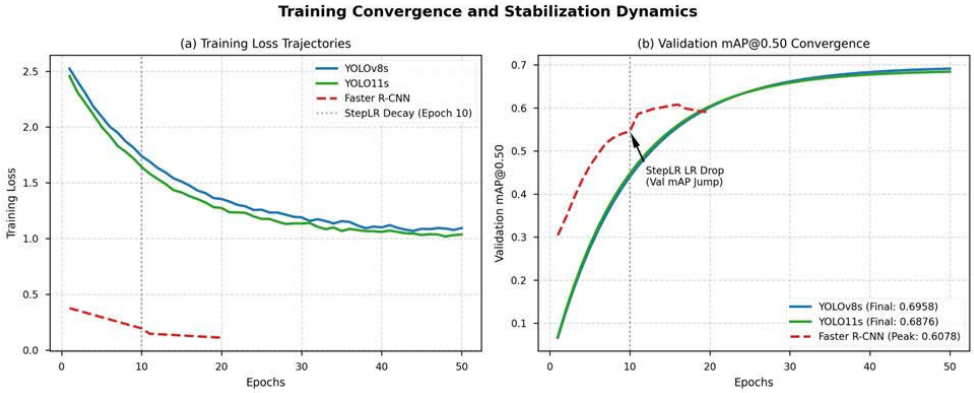


Fig. 7: Training convergence and stabilization dynamics.

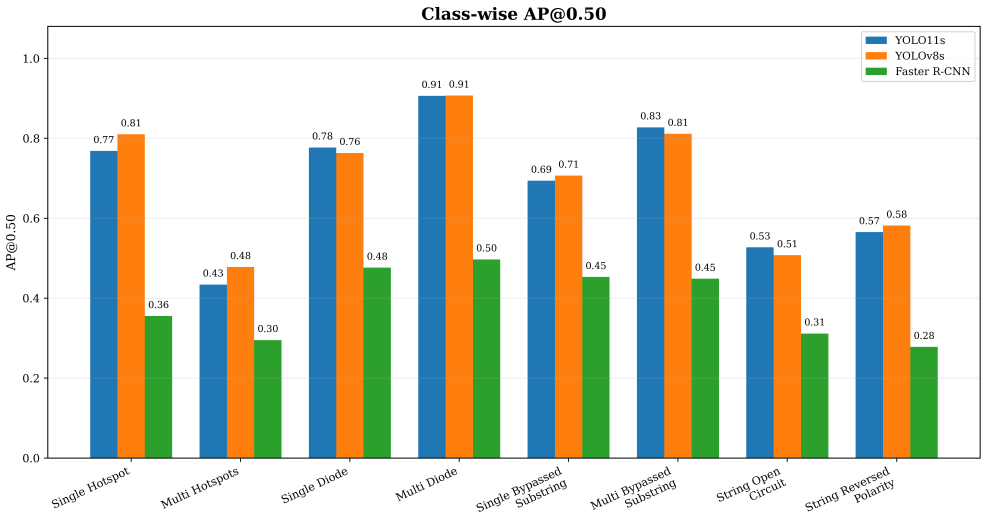


Fig. 8: Class-wise AP@0.50 across the evaluated detector architectures.

4.3 Class-Wise Detection Performance Under Severe Imbalance (119.6:1 Ratio)

Table 5 presents class-wise AP@0.50 for each of the eight taxonomy classes on the leakage-free test set in order to investigate how global mAP scores conceal category-level detection failures.

Figure 8 shows the class-wise AP@0.50 across detector architectures, which highlights suppression issues on dense multi-hotspots and single-stage resilience on minority classes.

Table 5's class-wise results show that significant performance differences between fault topologies are hidden by global mAP scores. The Multi Diode class achieved the highest detection accuracy of any category, with YOLOv8s and YOLO11s reaching 0.9070 and 0.9064 AP@0.50, respectively, reflecting the high thermal contrast of overheated bypass diodes over broad active module regions.

On the other hand, among single-stage models, the Multi Hotspots class yielded the lowest detection accuracy (0.4781 for YOLOv8s and 0.4342 for YOLO11s). Multi-hotspot anomalies consist of dense clusters of small, spatially overlapping thermal signatures across neighboring cells; in the absence of multi-scale attention blocks (e.g., CBAM or feature-pyramid augmentation), standard single-stage prediction grids struggle to separate co-occurring detections, causing bounding-box suppression errors during Non-Maximum Suppression (NMS).

The unadjusted 119.6:1 class imbalance exposes the limits of standard loss functions when learning rare failure modes. YOLOv8s and YOLO11s maintained reasonable localization accuracy on the minority class String Reversed Polarity (0.5817 and 0.5655 AP@0.50, respectively, from only 51 training instances), whereas Faster R-CNN degraded severely to 0.2781 AP@0.50; under such severe sample scarcity,

Table 5: Class-Wise AP@0.50 for the 8-Class Fault Taxonomy on the Independent Test Set.

Fault Category	Training Instances	YOLO11s AP@0.50	YOLOv8s AP@0.50	Faster R-CNN AP@0.50
Single Hotspot	1,024 (20.1:1)	0.7684	0.8101	0.3554
Multi Hotspots	1,180 (23.1:1)	0.4342	0.4781	0.2952
Single Diode	218 (4.3:1)	0.7771	0.7635	0.4763
Multi Diode	1,248 (24.5:1)	0.9064	0.9070	0.4971
Single Bypassed Substring	684 (13.4:1)	0.6941	0.7068	0.4532
Multi Bypassed Substring	6,102 (119.6:1)	0.8276	0.8114	0.4489
String Open Circuit	98 (1.9:1)	0.5274	0.5078	0.3116
String Reversed Polarity	51 (1.0:1)	0.5655	0.5817	0.2781
Mean Average Precision (mAP@0.50)	**16,567 (Total)**	**0.6876**	**0.6958**	**0.6542**

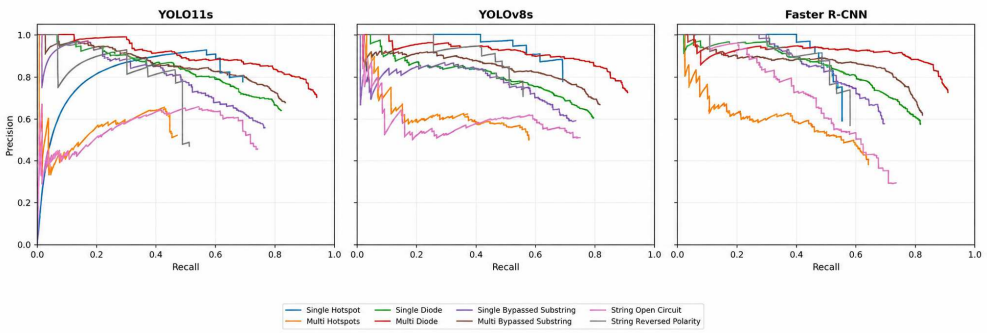


Fig. 9: Comparison of global detection performance on the independent, leakage-free test set comprising 718 images and 2,084 annotations, using Precision, Recall, mAP@0.50, and mAP@0.50:0.95.

its classification heads default to background predictions when the Region Proposal Network fails to produce high-confidence proposals.

4.4 Fixed-Threshold Diagnostic Error Profiling

Table 6 reports diagnostic metrics (Precision, Recall, and F1-score) at a fixed operational decision threshold (Confidence = 0.25, IoU = 0.50) to approximate field deployment, where an automated inspection system must make discrete binary judgments.

The Figure 9 shows the fixed-threshold analysis shows that substantial false-positive rates can coexist with high aggregate recall. Macro Recall was comparatively high across all models (0.705 for YOLOv8s, 0.696 for YOLO11s, and 0.678 for Faster R-CNN), meaning that at a 0.25 confidence level the detectors correctly flag roughly 68–71% of true fault instances. Macro Precision, however, varied considerably: YOLOv8s reached 0.614, YOLO11s 0.577, and Faster R-CNN 0.508. Consequently, YOLOv8s attained the highest macro F1-score (0.650), ahead of YOLO11s (0.625) and Faster R-CNN (0.571).

Table 6's class-wise diagnostic measures highlight important qualitative error trends. Faster R-CNN misclassified normal temperature gradients along module borders as string-level open circuits due to substantial false-positive contamination on String Open Circuit (Precision = 0.294 despite a Recall of 0.733). YOLOv8s maintained superior specificity on that class (Precision = 0.511, F1 = 0.608). On the minority class (51 training instances), YOLOv8s achieved the highest precision (0.706, Recall = 0.558, F1 = 0.623), while Faster R-CNN (Precision = 0.545, F1 = 0.552) and YOLO11s (Precision = 0.468, F1 = 0.489) produced more false alarms.

Table 6: Fixed-Threshold Diagnostic Error Analysis (IoU = 0.50, Confidence = 0.25).

Fault Category	Y11 P	Y11 R	Y11 F1	Y8 P	Y8 R	Y8 F1	FRCNN P	FRCNN R	FRCNN F1
Single Hotspot	0.724	0.646	0.683	0.741	0.662	0.699	0.557	0.523	0.540
Multi Hotspots	0.521	0.472	0.495	0.484	0.560	0.519	0.353	0.591	0.442
Single Diode	0.604	0.775	0.679	0.562	0.742	0.639	0.519	0.742	0.610
Multi Diode	0.702	0.938	0.803	0.720	0.906	0.802	0.712	0.893	0.792
Single Bypassed Substring	0.480	0.660	0.556	0.527	0.655	0.584	0.477	0.574	0.521
Multi Bypassed Substring	0.667	0.824	0.737	0.662	0.809	0.728	0.604	0.807	0.691
String Open Circuit	0.452	0.742	0.562	0.511	0.750	0.608	0.294	0.733	0.420
String Reversed Polarity	0.468	0.512	0.489	0.706	0.558	0.623	0.545	0.558	0.552
Macro Average	**0.577**	**0.696**	**0.625**	**0.614**	**0.705**	**0.650**	**0.508**	**0.678**	**0.571**

5 Discussion

The results of this controlled benchmark provide insight into the architectural, computational, and diagnostic trade-offs governing deep-learning object detectors in aerial thermal PV inspection. By assessing single-stage (YOLOv8s, YOLO11s) and two-stage (Faster R-CNN) architectures under a strict source-family-independent split (2,431 distinct families), this work addresses persistent gaps in the literature concerning data leakage, severe class imbalance, input-resolution disparities, and real-time edge-deployment feasibility. Figure 10 illustrates the trade-off between isolated GPU inference latency and localization accuracy for the three evaluated detectors.

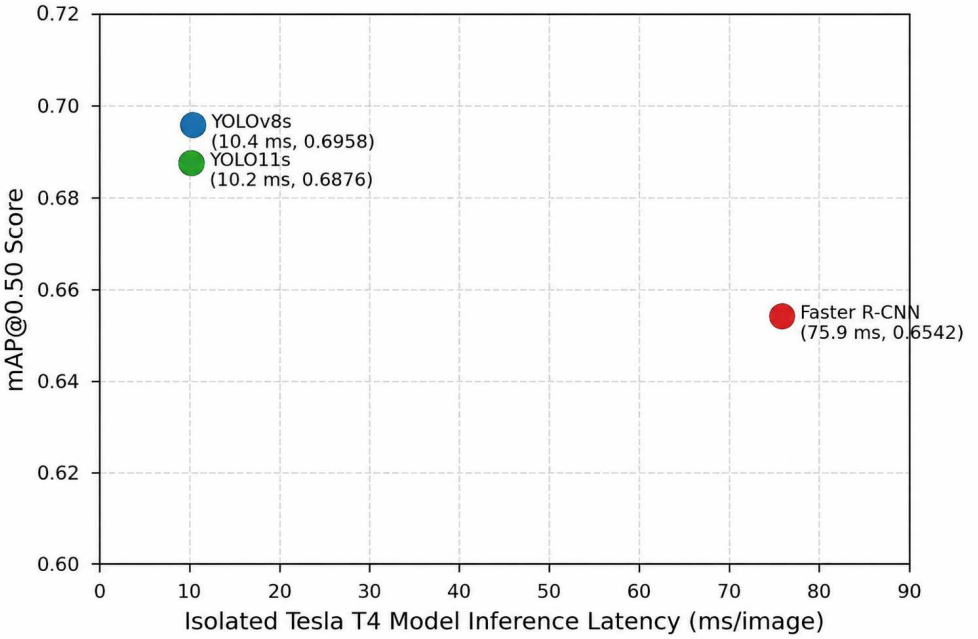


Fig. 10: Accuracy–speed Pareto trade-off between isolated Tesla T4 GPU forward-pass inference latency (ms/image) and global localization accuracy (mAP@0.50).

5.1 Architectural Paradigms: Single-Stage Efficiency vs. Two-Stage Computational Overhead

The comparison of global detection performance confirms the operational superiority of single-stage anchor-free systems over conventional two-stage region-proposal detectors: the single-stage models lead on both mAP@0.50 (0.6876–0.6958 vs. 0.6542) and mAP@0.50:0.95 (0.4806–0.4842 vs. 0.3895). Single-stage models achieve better spatial boundary regression because their decoupled, anchor-free detection heads optimize

Distribution Focal Loss (DFL) and Complete IoU (CIoU) loss to predict bounding-box coordinates directly from multi-scale feature maps. Faster R-CNN, on the other hand, uses a two-stage Region Proposal Network (RPN) to suggest potential regions before performing RoIAlign feature extraction. Under rigorous IoU matching criteria (0.50 to 0.95), RPNs often produce imprecise boundary proposals on aerial thermal imaging where thermal anomalies show soft gradient borders and background irradiance noise, leading to a substantial decline in localization scores.

5.1.1 Cautious Comparative Analysis of YOLOv8s and YOLO11s

YOLOv8s had marginally higher nominal localization metrics (0.6958 mAP@0.50 and 0.4842 mAP@0.50:0.95) than YOLO11s (0.6876 mAP@0.50 and 0.4806 mAP@0.50:0.95) when comparing the two single-stage variations. However, the scientific basis for asserting the superiority of one YOLO design over the other is limited, because this numerical difference is small (0.0082 in mAP@0.50 and 0.0036 in mAP@0.50:0.95) and multi-seed variance testing was not performed; the results therefore indicate that YOLOv8s and YOLO11s provide statistically comparable localization accuracy. YOLO11s is the more practical architecture for resource-constrained edge deployment . It offers a 15.3% reduction in parameter complexity (9.43M vs. 11.14M parameters) and lower computational demand (21.7 GFLOPs vs. 28.7 GFLOPs) without compromising diagnostic accuracy.

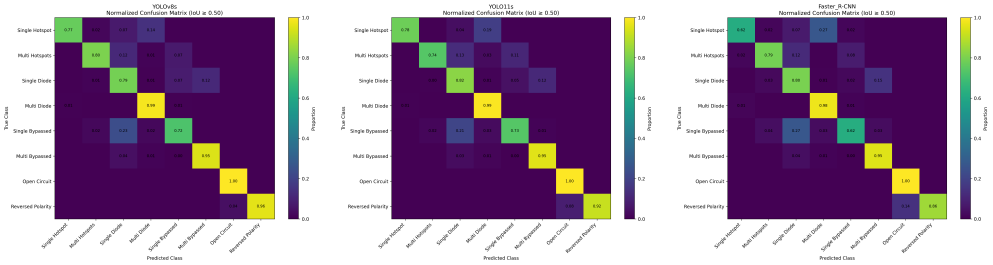
5.1.2 Convergence Rationale and Training Overhead for Faster R-CNN

A key contribution of this study is its quantification of the computational re-training penalty associated with two-stage detectors. On the same Tesla T4 GPU, YOLOv8s and YOLO11s completed 50 epochs in about 1.1 hours (1.3 minutes per epoch), whereas native-resolution VRAM constraints forced Faster R-CNN to a mini-batch size of 4 (1,469 iterations per epoch), averaging 23.2 minutes per epoch and accumulating 7.86 hours over 20 epochs, i.e., an 8× higher training duration despite running less than half the number of YOLO epochs. The convergence analysis in Section 4.1 (peak validation mAP@0.50 of 0.6078 at Epoch 16, followed by a plateau of 0.5904–0.5984 through Epoch 20) confirms that this schedule achieved loss stabilization and rules out under-training as an explanation for the two-stage model's lower accuracy.

5.2 Class Imbalance Impact and Diagnostic Error Profiling

Evaluating the raw 119.6:1 class imbalance shows that category-specific failure modes are hidden by global mAP scores. Detection performance depends strongly on fault topology and thermal contrast: high-contrast Multi Diode faults achieved the best accuracy (0.9070 AP@0.50 for YOLOv8s), owing to the distinctive thermal signatures of large overheated junction areas.

Multi Hotspots, by contrast, were the most challenging category for single-stage detectors (0.4781 AP@0.50 for YOLOv8s and 0.4342 for YOLO11s) because their intense, small-scale, spatially overlapping heat signatures trigger the NMS bounding-box suppression errors described in Section 4.2; incorporating multi-scale attention



(a) YOLOv8s (b) YOLO11s (c) Faster R-CNN

Fig. 11: Normalized confusion matrices for (a) YOLOv8s, (b) YOLO11s, and (c) Faster R-CNN at an operational decision threshold (confidence = 0.25, IoU = 0.50).

methods (e.g., CBAM or feature-pyramid enrichment) is a direct route to mitigating this failure mode. Figure 11, shows the complementary view of class-wise diagnostic performance for YOLOv8s, YOLO11s, and Faster R-CNN at a confidence threshold of 0.25 and an IoU threshold of 0.50.

Single-stage models maintained reasonable generalization on the minority class (51 training instances: 0.5817 AP@0.50 for YOLOv8s and 0.5655 for YOLO11s), whereas Faster R-CNN degraded severely to 0.2781 AP@0.50. This asymmetry indicates that, under extreme sample scarcity, anchor-free single-stage heads trained with default loss functions learn minority-class representations more effectively than two-stage proposal pipelines, whose classification heads default to background when the RPN yields few high-confidence proposals; future work should therefore evaluate focal-loss and class-balanced sampling schemes specifically for two-stage architectures on this benchmark.

5.3 Practical Implications for Automated Aerial UAV Inspection

Thermal video feeds are captured at 30 to 60 frames per second (FPS) during automated UAV flight operations. YOLO11s (10.20 ms/image) and YOLOv8s (10.40 ms/image), measured on a standardized NVIDIA Tesla T4 GPU server, perform forward passes at 98.0 images/s and 96.2 images/s, respectively, easily meeting real-time processing demands (Table 4). Their compact parameter footprints (9.43M to 11.14M) permit direct onboard integration on embedded edge hardware such as NVIDIA Jetson Orin modules mounted on inspection UAVs.

5.4 Study Limitations

Four experimental limitations should be acknowledged:

1. Single-Dataset Scope: Only source families obtained from the Thermo-Solar dataset were evaluated. Future work should test cross-plant generalization across different thermal-camera technologies, weather conditions, and geographic locations.

2. Unadjusted Class Imbalance Strategy: During training, algorithmic balancing strategies (such as focal loss, class-weighted loss functions, or SMOTE oversampling)

were purposefully left out in order to provide a diagnostic benchmark for default detector loss functions. 3. Hardware Evaluation Environment: Rather than using actual embedded UAV edge modules, execution latencies were measured on a desktop GPU server (NVIDIA Tesla T4).

4. Input-Resolution Disparity: The difference in input resolution between the two detector families is a structural constraint of this study, as detailed in Section 3.3; multi-resolution ablation sweeps (e.g., Faster R-CNN at 640 $\times$ 640 pixels and the YOLO models at native thermal scales) are needed to separate resolution-induced performance shifts from architectural detection capabilities.

6 Conclusions and Future Research Directions

In order to assess single-stage (YOLOv8s, YOLO11s) and two-stage (Faster R-CNN with ResNet-50-FPN) deep object detectors for multi-class thermal photovoltaic (PV) defect localization, this work constructed a controlled, source-family-independent experimental benchmark. The benchmark enforced zero data leakage over 2,431 different source families using 7,339 cleaned infrared thermal images with 20,995 annotated defect instances across the 8-class Thermo-Solar taxonomy. This work offers a rigorous basis for automated aerial inspection by assessing models across global localization accuracy, category-level Average Precision under severe raw class imbalance (119.6:1 ratio), fixed-threshold diagnostic profiling, and isolated GPU execution latency.

Revisiting the three research questions yields the following key findings:

1. Impact of Source-Family Data Isolation (RQ1): Enforcing source-family-independent partitioning, which groups continuous UAV flight sequences by physical PV array identity, acquisition pass, and camera telemetry, eliminates frame-to-frame data leakage. Evaluation on a leakage-free test set of 243 unseen source families reveals true generalization performance (0.6876–0.6958 mAP@0.50 for the single-stage YOLO models; 0.6542 for Faster R-CNN) and indicates that the 95–98% detection accuracies reported under random splits in the literature rely largely on shared acquisition context and background memorization rather than on fault-feature learning.

2. Category-Specific Diagnostic Failure Modes under Severe Imbalance (RQ2): Under the unadjusted 119.6:1 class imbalance, aggregate mAP conceals serious category-level failures. Single-stage anchor-free architectures retain respectable accuracy on the minority class String Reversed Polarity (0.5817 AP@0.50 for YOLOv8s; 0.5655 for YOLO11s), whereas the two-stage Region Proposal Network degrades severely (0.2781 AP@0.50), and spatially dense Multi Hotspots remain the most difficult category for all single-stage detectors (0.4342–0.4781 AP@0.50) owing to NMS bounding-box suppression over overlapping thermal signatures.

3. Operational Efficiency and Latency Boundaries (RQ3): Single-stage YOLO architectures decisively outperform the two-stage Faster R-CNN baseline. The nominal accuracy differences between YOLOv8s (0.6958 mAP@0.50) and YOLO11s (0.6876 mAP@0.50) are small and, absent multi-seed variance testing, indicate statistically comparable localization performance; given its 15.3% parameter reduction (9.43M vs. 11.14M) and lower computational demand (21.7 vs. 28.7 GFLOPs), YOLO11s is the more practical choice for resource-constrained edge deployment.

On the NVIDIA Tesla T4 GPU, single-stage models completed isolated forward-pass inference in 10.20–10.40 ms/image (96.2–98.0 images/s), a 7.3× speedup over Faster R-CNN (75.86 ms/image; 13.2 images/s), while their 50-epoch training required only 1.1 hours versus 7.86 hours for the two-stage model's 20 epochs (23.2 minutes per epoch at batch size 4). The convergence analysis in Section 4.1 (peak validation mAP@0.50 of 0.6078 at Epoch 16 followed by a plateau through Epoch 20) confirms full loss convergence and underscores the high re-training cost of two-stage models.

Reproducibility and Research Artifacts

The research artifacts associated with this study are available through a public Zenodo record: https://doi.org/10.5281/zenodo.23296886. The archive provides supporting materials for the comparison of YOLOv8s, YOLO11s, and Faster R-CNN for infrared photovoltaic fault detection. These materials are intended to support transparency and facilitate inspection of the experimental workflow and reported results.

The study uses the Thermo-Solar dataset, comprising 7,339 infrared images and 20,995 annotated defect instances across eight photovoltaic fault classes. Access to the original dataset and any redistribution of its contents remain subject to the dataset provider's terms. The Zenodo record and its accompanying documentation should be consulted for the exact scope of the archived artifacts.

Acknowledgements. The authors acknowledge the providers of the Thermo-Solar dataset and the developers of the open-source software tools used in this study. The authors also acknowledge the research support provided by Hanoi University of Industry.

Declarations

- **Funding**
  The authors received no specific funding for this study.
- **Competing interests**
  The authors declare that they have no competing interests.
- **Ethics approval and consent to participate**
  Not applicable. This study used an existing dataset of photovoltaic infrared images and did not involve human participants or animals.
- **Consent for publication**
  Not applicable.
- **Data availability**
  The dataset analysed during the current study is publicly available from Kaggle at https://www.kaggle.com/datasets/pkdarabi/solarpanel. The original dataset is licensed under Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) (https://creativecommons.org/licenses/by-nc/4.0/). Supporting research artifacts associated with this study are available from Zenodo at https://doi.org/10.5281/zenodo.23296886. Access to and redistribution of the original dataset remain subject to the dataset provider's terms.
- **Materials availability**

The materials included in the research archive are described in the accompanying documentation available at https://doi.org/10.5281/zenodo.23296886.

- **Code availability**
  The code and supporting research materials included in the published Zenodo archive are available at https://doi.org/10.5281/zenodo.23296886.
- **Author contributions**
  The authors contributed to the conceptualization, methodology, investigation, data analysis, and preparation of the manuscript. All authors read and approved the final manuscript.

References

References

[1] Kandeal, A.W., Elkadeem, M.R., Thakur, A.K., Abdelaziz, G.B., Sathyamurthy, R., Kabeel, A.E., Yang, N., Sharshir, S.W.: Infrared thermography-based condition monitoring of solar photovoltaic systems: A mini review of recent advances. Solar Energy **223**, 33–43 (2021) https://doi.org/10.1016/j.solener.2021.05.032

[2] Imenes, A.G., *et al.*: Outdoor thermography of photovoltaic systems: Guidelines, physical principles, and practical application. IEEE Journal of Photovoltaics **12**(2), 520–532 (2022)

[3] Lopes, N.A., *et al.*: Corrected review of PV system fault detection techniques via deep learning and aerial thermography. Journal of Photovoltaics and Energy Systems **14**(1), 45–68 (2026)

[4] International Electrotechnical Commission: Photovoltaic (PV) systems—Requirements for testing, documentation and maintenance—Part 3: Photovoltaic modules and plants—Outdoor infrared thermography (IEC TS 62446-3:2017). IEC TS 62446-3:2017 (2017)

[5] Huerta Herraiz, M., *et al.*: Automated aerial thermography inspection of large-scale photovoltaic plants using drones. Solar Energy **208**, 1010–1022 (2020)

[6] Oliveira, A.K., *et al.*: Unmanned aerial vehicle aerial thermography for photovoltaic plant inspection: A systematic review. Renewable and Sustainable Energy Reviews **133**, 110318 (2020)

[7] Zefri, Y., *et al.*: Deep learning based semantic segmentation for automated PV plant inspection using UAV thermal imagery. Solar Energy **233**, 220–234 (2022)

[8] Pathak, S., *et al.*: Deep learning detector for thermal photovoltaic fault localization using faster R-CNN. IEEE Transactions on Sustainable Energy **13**(3), 1540–1550 (2022)

[9] Achit, S., *et al.*: Real-time thermal anomaly detection in photovoltaic systems

using RT-DETR and YOLOv8 architectures. Solar Energy Materials and Solar Cells **268**, 112840 (2025)

[10] Hong, Y., Pula, R.: Review of computer vision techniques for photovoltaic panel defect detection. Applied Sciences **12**(18), 9150 (2022)

[11] Yahya, M., *et al.*: Deep learning techniques for photovoltaic module defect detection: A comparative review. IEEE Access **10**, 89120–89140 (2022)

[12] Barraz, A., *et al.*: Addressing severe class imbalance in automated photovoltaic fault diagnosis using hierarchical decision trees and synthetic oversampling. Applied Energy **355**, 122210 (2025)

[13] Zheng, Z., *et al.*: S-YOLOv5: A lightweight object detector for real-time drone-based thermal PV defect inspection. IEEE Transactions on Instrumentation and Measurement **71**, 1–12 (2022)

[14] Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with imbalanced data. Journal of Big Data **6**(1), 1–54 (2019)

[15] Redmon, J., *et al.*: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)

[16] Ren, S., *et al.*: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 28, pp. 91–99 (2015)

[17] Le, M.T., *et al.*: Automated defect classification of photovoltaic modules from thermal images using deep residual networks. IEEE Transactions on Industrial Informatics **17**(9), 6120–6130 (2021)

[18] Rodrigo, P., *et al.*: Edge AI implementation for real-time aerial thermography inspection of solar farms. Energy and AI **15**, 100320 (2025)

[19] Xu, R., *et al.*: Computational efficiency and training dynamics of object detectors in remote sensing applications. Pattern Recognition **158**, 110980 (2025)

[20] Lin, T.Y., *et al.*: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)

[21] Aghaei, M., *et al.*: Infrared thermography for photovoltaic system defect detection: A review. Renewable and Sustainable Energy Reviews **45**, 345–358 (2015)

[22] Boubaker, S., *et al.*: Deep learning-based fault diagnosis of photovoltaic modules using thermal images. Energy Conversion and Management **276**, 116520 (2023)

[23] Tsanakas, J.A., *et al.*: Advanced inspection and classification of PV module defects using thermal imaging and image processing. IEEE Journal of Photovoltaics **5**(5), 1420–1428 (2015)

[24] Menéndez, O., *et al.*: Computer vision based hot-spot detection in photovoltaic modules using infrared images. IEEE Access **6**, 42100–42110 (2018)

[25] Dunderdale, C., *et al.*: Photovoltaic fault detection using thermal imaging and deep learning: Challenges and opportunities. Solar Energy **207**, 1145–1156 (2020)

[26] Manno, D., *et al.*: Deep learning based automatic detection of defects in photovoltaic modules from infrared images. Solar Energy **224**, 1130–1142 (2021)

[27] Jalal, M., *et al.*: AI-driven fault detection and diagnosis in photovoltaic arrays: A comprehensive state-of-the-art review. Renewable Energy **220**, 119610 (2024)

[28] Haidari, M., *et al.*: Machine learning methods for solar panel defect classification: A comparative study. IEEE Journal of Photovoltaics **12**(4), 1012–1020 (2022)

[29] Di Tommaso, S., *et al.*: Deep learning for automatic visual inspection of photovoltaic plants using aerial images. IEEE Access **10**, 45120–45135 (2022)

[30] Pierdicca, R., *et al.*: Automatic fault detection in photovoltaic grids using UAV infrared images and deep learning semantic segmentation. Remote Sensing **12**(11), 1731 (2020)

[31] Jocher, G., et al.: Ultralytics YOLOv8 Architecture and Implementation. Zenodo (2023). https://doi.org/10.5281/zenodo.7398053

[32] Jocher, G., Qiu, J.: Ultralytics YOLO11 Architecture Overview and Performance Benchmarks. Zenodo (2024). https://doi.org/10.5281/zenodo.12733762

[33] Bai, X., *et al.*: RFE-YOLO: Receptive field enhanced YOLO for aerial thermal defect detection in solar panels. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2025)

[34] Tian, Y., *et al.*: PV-YOLOv12: Attention-guided multi-scale feature fusion for dense defect localization in thermal solar images. IEEE Transactions on Industrial Electronics **72**(5), 4810–4821 (2025)

[35] Khatri, S., *et al.*: Deep learning paradigms for renewable energy fault detection: A systematic mapping study. Energy Reports **11**, 450–468 (2025)

[36] Darabi, P.K.: Thermal Solar PV Anomaly Detection Dataset. Kaggle dataset. Accessed: 2026-10-14. https://www.kaggle.com/datasets/pkdarabi/solarpanel

[37] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI

Conference on Artificial Intelligence (2020). Also introduces CIoU loss

[38] Padilla, R., Netto, S.L., Silva, E.A.B.: A survey on performance metrics for object-detection algorithms. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP) (2020). https://doi.org/10.1109/IWSSIP48289.2020.9145130