

Article

Physics-Guided Deep Learning and Geospatial Aggregation for Photovoltaic Panel Detection and Duplicate-Free Counting in Unmanned Aerial Vehicle Imagery

Sixin Zhu *, Kunping Liang, Xu Zhao, Yijie Hu and Hongqian He

School of Earth Sciences and Engineering, North China University of Water Resources and Electric Power, No. 136 Jinshui East Road, Zhengzhou 450046, China; kunpingliang2025@163.com (K.L.); xuzhao139@gmail.com (X.Z.); z20251020244@stu.ncwu.edu.cn (Y.H.); hehongqian@ncwu.edu.cn (H.H.)

* Correspondence: zhusixin@ncwu.edu.cn

Featured Application

The proposed framework can be used for automated inventory and inspection of utility-scale photovoltaic plants. It detects individual panels in overlapping unmanned aerial vehicle images and converts repeated frame-level observations into unique georeferenced assets. The resulting panel locations and counts can support asset indexing, inspection planning, maintenance prioritization, and comparison of repeated surveys without the computational burden and alignment errors associated with constructing a full image mosaic.

Abstract

Automated inspection of utility-scale photovoltaic plants requires reliable detection and inventory of individual panels from overlapping aerial images. We developed a physics-guided deep learning and geospatial aggregation framework for panel detection and duplicate-free plant-level counting. Scene-targeted augmentation introduces bounded specular-reflection, occlusion, perspective, and boundary-ambiguity perturbations into training images. A You Only Look Once version 11 (YOLO11)-based detector combines adaptive input resolution, reflection-oriented channel attention, and a decoupled feature pyramid network. Frame-level detections are projected into a common geographic coordinate system using unmanned aerial vehicle (UAV) pose and camera geometry, and spatial clustering merges repeated observations without image stitching. On 4200 UAV images containing approximately 110,000 annotated panel instances, the complete detector achieved 74.3% average precision at an intersection-over-union threshold of 0.5 (AP@0.5) and 62.9% AP@0.5:0.95 across five runs. Geospatial aggregation achieved 99.2% plant-level counting accuracy, compared with 75.1% for image stitching. On an independent held-out desert-site dataset containing approximately 45,000 panel instances, the Site-A-trained model was applied without retuning and retained 72.8% AP@0.5 and 98.4% plant-level counting accuracy. Component, attention, regional-counting, layout, cross-site, and runtime experiments further evaluated the framework. Final-pass throughput was 105.6 frames/s; a separate two-pass runtime experiment reported 15.1 ms/image for the complete processing chain, equivalent to 66.2 frames/s. The results support integrated panel detection and inventory under the evaluated imaging and survey conditions.



Academic Editor: Thomas Lindner

Received: 28 August 2026

Revised: 23 September 2026

Accepted: 28 September 2026

Published: 30 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: photovoltaic inspection; unmanned aerial vehicle imagery; small-object detection; physics-guided deep learning; geospatial aggregation; asset inventory

1. Introduction

Utility-scale photovoltaic (PV) plants contain large numbers of modules whose condition and inventory must be monitored throughout operation and maintenance. Manual inspection and counting are labor-intensive, while aerial imaging provides a practical way to acquire detailed information over large sites [1–4]. In particular, unmanned aerial vehicles (UAVs) offer flexible deployment, high spatial resolution, and repeatable acquisition, making them suitable for automated module detection, defect screening, and asset inventory [3,5–9].

However, reliable panel inventory from UAV image sequences is more difficult than single-frame detection. Variations in flight altitude, terrain relief, and viewing geometry change apparent target size and perspective; glass surfaces produce strong view- and illumination-dependent specular reflection; and dense arrays introduce occlusion and boundary ambiguity. In addition, standard flight plans intentionally use high image overlap, so the same physical module is observed in several frames. A practical inspection system must therefore address both difficult image-level detection and duplicate removal across the full survey [6,10,11].

Existing studies have improved photovoltaic panel or defect detection using deep learning, thermal and visible imagery, multi-scale feature fusion, and attention mechanisms [5–9,12–22]. Other work has explored photovoltaic mapping from high-resolution imagery or geospatial registration of detected modules [6,23–26]. These lines of research solve important parts of the problem, but an operational plant inventory requires them to work together: detection errors affect the reliability of downstream mapping, while accurate frame-level detections can still produce severe over-counting when overlapping observations are summed directly.

We propose a photovoltaic panel detection and geospatial aggregation system, referred to as PVGS, to address this engineering problem. The system combines physics-guided image augmentation, an improved You Only Look Once version 11 (YOLO11) detector, and pose-assisted geographic aggregation. Its objective is to improve detection under challenging photovoltaic imaging conditions and convert repeated observations into unique physical assets. The main contributions are as follows:

- (1) PV-PhysAug is a physics-guided augmentation strategy that introduces bounded reflection, occlusion, perspective, and boundary perturbations into UAV training images. These operations represent illumination- and geometry-related imaging variability while preserving the acquired validation and test data.
- (2) A module-level detector built on YOLO11 incorporates adaptive input resolution, a reflection-aware module (RAM), and a decoupled feature pyramid network (DFPN). These components preserve small-object detail and improve robustness to specular interference and scale variation in UAV imagery.
- (3) A pose-assisted geospatial aggregation method combines frame-level detections with UAV position and orientation data, camera geometry, and a surface constraint. Spatial clustering merges repeated observations in a common geographic frame, producing unique panel locations and plant-level counts without an intermediate stitched orthomosaic.

We evaluated image-level detection and plant-level aggregation at a mountainous photovoltaic plant covering more than 4 km². The detector achieved an average precision (AP) of 74.3% at an intersection-over-union (IoU) threshold of 0.5, and geographic aggregation achieved 99.2% plant-level counting accuracy. A separate held-out desert photovoltaic site containing approximately 45,000 panel instances was then evaluated without retuning the Site-A model, providing an independent cross-site test. Repeated-run ablation, isolated component tests, attention comparisons, regional inventory checks, layout-dependent eval-

uation, cross-site validation, and two-pass timing experiments examine complementary aspects of the workflow.

2. Related Work

Automated photovoltaic inspection combines fine-scale visual detection with plant-level asset management. Previous research can be grouped into three closely related directions: photovoltaic panel and defect detection from aerial, thermal, or electroluminescence imagery [5–9,13–19], small-object detection and feature-fusion methods for complex scenes [20–22,27–29], and photovoltaic mapping or geospatial labeling for spatially explicit inventories [6,23–26]. The present study connects these directions by combining a panel detector with cross-frame geographic aggregation.

2.1. Single-Frame Detection Under Challenging Imaging Conditions

Challenges in detecting PV modules from individual aerial frames arise primarily from the physical imaging conditions encountered during UAV inspection. As shown in Figure 1, three factors are especially important:



Figure 1. Imaging conditions in the acquired UAV data. (a) Perspective distortion; (b) strong specular reflection; (c) array overlap and boundary ambiguity.

First, dramatic scale variation and perspective distortion (Figure 1a), caused by high-altitude nadir viewpoints, create a difficult small-object detection scenario and increase localization uncertainty [8,12]. Second, strong specular reflections on PV module surfaces under certain illumination (Figure 1b) act as structured optical interference, overwhelming intrinsic texture features and reducing foreground-background contrast [5,13]. Finally, dense installation, occlusion, and blurred boundaries (Figure 1c) cause feature entanglement and instance ambiguity, complicating the separation of adjacent modules and increasing missed or duplicate detections [7,8].

To address these image-interpretation challenges, object-detection research has increasingly used enhanced feature pyramids, multi-level feature fusion, and attention mechanisms to preserve small-target information and suppress background interference [16,20–22]. Previous studies have also shown that attention-guided enhancement and multi-scale feature fusion can improve recognition performance in complex image-analysis tasks [27–29]. Recent UAV-oriented detectors further emphasize efficient transformer architectures and stronger multi-scale feature extraction for small targets [30,31]. Most of these advances remain image-centric: they improve the representation of small or difficult targets without explicitly addressing the repeated-observation problem at plant scale.

Data augmentation is especially relevant when difficult operating conditions are underrepresented even in a numerically large training set. Generic geometric and photometric

augmentation can improve coverage of the empirical data distribution, but domain-targeted augmentation should remain bounded so that it does not create implausible samples. The amount of data needed for stable learning is task-dependent; theoretical learning-curve work has shown that the reduction in excess classification error depends on training-set size and feature dimensionality [32]. Because the present task is object detection rather than parametric Bayes classification, we do not transfer a theoretical minimum sample size directly. Instead, the 4200-image dataset is combined with geographically independent splitting, repeated training runs, and bounded augmentation, while the scope and limitations of the primary evaluation are discussed explicitly.

2.2. Cross-Frame Aggregation and Plant-Level Counting

Even with improved per-frame detection accuracy, obtaining a unique plant-level count requires resolving repeated observations caused by highly overlapping UAV images. Naïve summation of per-frame detections leads to overcounting because the same physical module can appear in multiple frames [6,8]. Figure 2 illustrates three sequential UAV frames and their corresponding detections, showing how the same panels are repeatedly observed in overlapping images. A globally consistent reference frame is therefore required for duplicate removal.

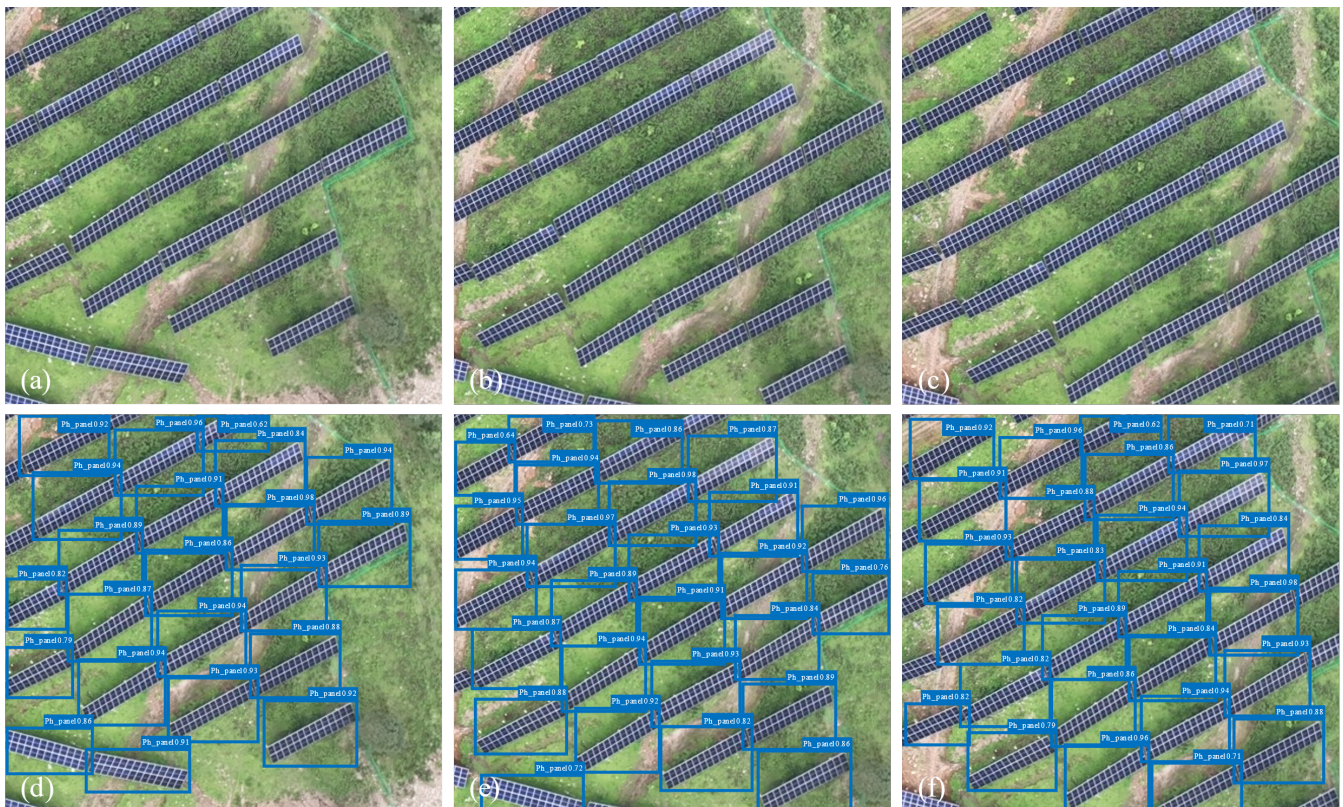


Figure 2. Sequential UAV frames and corresponding detections. (a–c) Original frames 1–3; (d–f) frame-level detection results for the respective frames.

Existing global counting strategies fall into two main categories:

The first category relies on image-stitching-based visual mapping. These methods generate a panoramic orthophoto by aligning and blending input images and then perform detection on the stitched result. Although visually intuitive, stitching is computationally expensive and vulnerable to alignment errors in repetitive PV-array textures. Such geometric inconsistencies can propagate directly into the final count.

The second category uses geographic registration and sensor fusion to associate image observations with physical locations [6]. In this strategy, each detection is projected into a common world coordinate system using UAV pose, camera geometry, and an elevation or surface model, after which spatial clustering can be used to merge repeated observations. Compared with a stitched-image workflow, this approach preserves an explicit geographic reference for each detected asset and avoids propagating local image-matching errors through a full mosaic. Its accuracy, however, depends on georeferencing precision, camera calibration, terrain representation, and the quality of the upstream detections.

Recent studies similarly show that small-object detection in UAV imagery benefits from multi-scale feature interaction and adaptive fusion [22], consistent with broader findings on bidirectional feature-pyramid information flow and progressive multi-level fusion [20,21]. These advances motivate stronger image-level representations, but they do not by themselves resolve the repeated-observation problem that arises when an aerial inspection survey must be converted from a collection of overlapping frames into a unique set of georeferenced physical assets.

2.3. Research Gap and Problem Formulation

The main research gap is the disconnect between image-space robustness and plant-level object identity. Mapping studies can delineate photovoltaic facilities over broad areas [24–26], and UAV inspection methods can identify panels or defects in individual images [5–9], but a module-level plant inventory requires both reliable detection under scene-specific radiometric and geometric disturbances and a mechanism for reconciling repeated observations of the same physical asset. Treating these stages independently can allow local detection errors to propagate into the final inventory or leave accurate detections vulnerable to systematic over-counting.

This study therefore links the two levels within a single UAV inspection workflow. Physics-guided augmentation and scale-aware feature extraction improve image-space robustness, while pose-constrained geographic projection and spatial clustering enforce cross-frame consistency. The intended output is a unique georeferenced panel inventory in which each cluster represents one physical module and can be integrated with asset-management systems or repeated inspection campaigns.

3. Materials and Methods

To address scale variation, specular reflection, occlusion, and repeated observations in UAV-acquired photovoltaic (PV) imagery, this study develops an integrated inspection workflow that couples image-space panel detection with geospatial aggregation. Figure 3 shows the overall architecture of the proposed PVGS system.

The framework contains three connected stages. During training, selected PV-PhysAug transformations are applied sequentially when the augmentation trigger is active (Figure 3a). During inference, a preliminary detector pass estimates scene density and selects the input resolution. The final pass then extracts multi-scale features, applies RAM and DFPN, and produces frame-level detections (Figure 3b). Geographic projection and spatial clustering combine these detections with UAV pose and camera geometry to obtain unique panel locations and a plant-level count (Figure 3c).

The enhanced YOLO11 detector addresses scale variation, specular interference, and multi-scale localization. RAM provides channel attention based on localized high-activation descriptors and is trained jointly with reflection-targeted image augmentation.

For plant-level counting, we combine frame-level detections with UAV position and orientation system (POS) data. Each detection is back-projected through the calibrated camera model and intersected with the selected surface representation. The reported

plant-level experiment uses a local horizontal plane within each survey block. Section 3.3 specifies this implementation together with the general surface-intersection equations. The projected points are clustered with DBSCAN to merge repeated observations and retain georeferenced panel locations without image stitching.

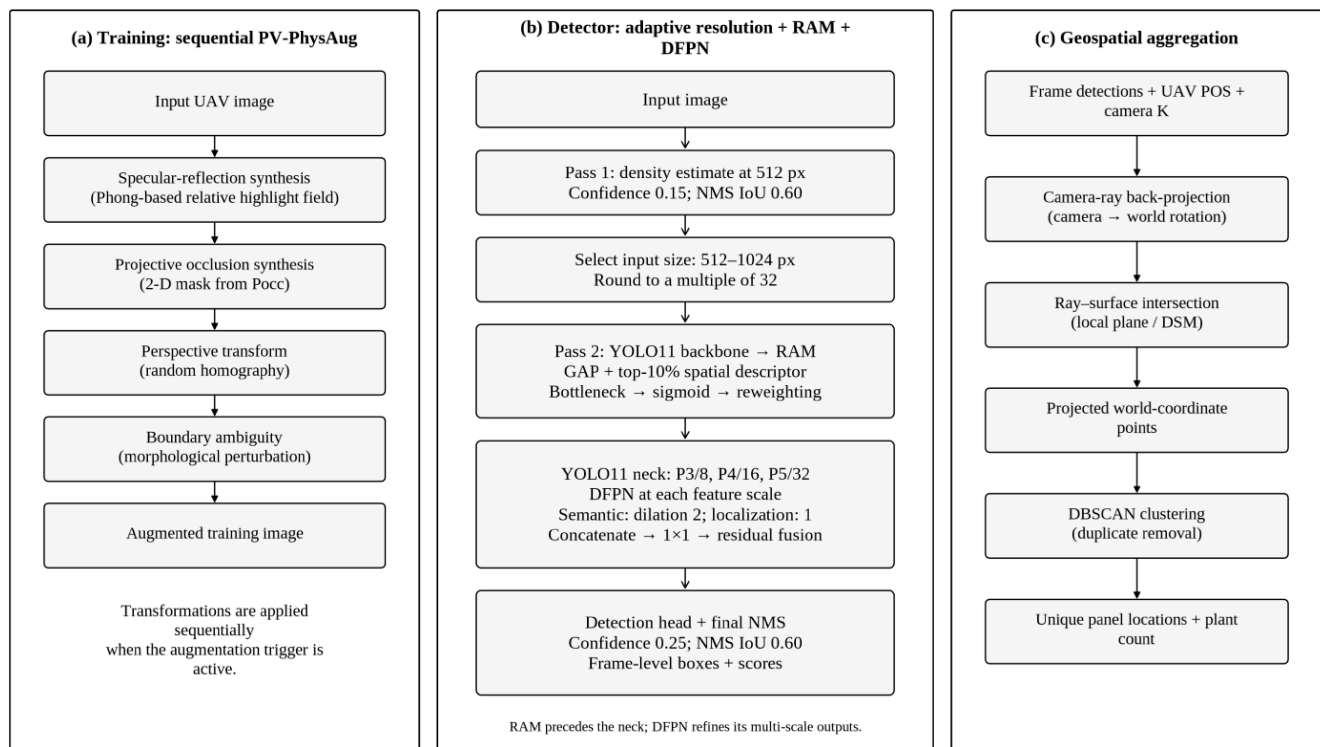


Figure 3. Architecture of PVGS. (a) Sequential PV-PhysAug training pipeline; (b) two-pass adaptive-resolution YOLO11-L detector with RAM and DFPN; (c) camera-to-world projection, ray-surface intersection, DBSCAN clustering, and singleton handling. POS: position and orientation system; DSM: digital surface model; DBSCAN: density-based spatial clustering of applications with noise; NMS: non-maximum suppression.

3.1. Physics-Guided Augmentation for Photovoltaic Imagery

The robustness of supervised image models is strongly influenced by the diversity and representativeness of the training samples. The UAV data already contain naturally occurring reflection, occlusion, perspective variation, and boundary ambiguity; however, severe combinations of these effects are relatively sparse. PV-PhysAug is therefore applied only to the training partition to increase coverage of these underrepresented nuisance conditions. An image enters the augmentation pipeline with probability $p_{\text{aug}} = 0.75$. Conditional on that trigger, reflection, occlusion, perspective, and boundary perturbations are sampled independently with probabilities 0.70, 0.50, 0.50, and 0.40, respectively, and the selected transformations are applied sequentially in the order shown in Figure 3a. Validation and test images are never augmented.

Specular reflection constitutes one of the most challenging sources of interference in photovoltaic panel detection. The implementation uses the simplified Phong formulation to compute relative highlight responses for training-image augmentation. For each annotated panel box, a locally planar normal is assigned and independently tilted by an angle sampled from -15° to 15° about the camera x and y axes. The solar azimuth is sampled uniformly from 0° to 360° and the solar elevation from 25° to 75° . The illumination direction is global within one augmented image, whereas the viewing direction varies spatially and is obtained from normalized pinhole rays across the image plane. The reflected direction follows the standard mirror relation $\mathbf{r} = 2(\mathbf{n} \cdot \mathbf{l})\mathbf{n} - \mathbf{l}$.

Equation (1) is used to compute the relative diffuse-plus-specular response. Only the specular excess above the diffuse baseline is converted to an additive highlight map R . Because the RGB images acquired with the DJI Mavic 3 Pro UAV (DJI, Shenzhen, China) are not radiometrically calibrated, sampled illuminance is not converted directly from lux to digital numbers. Instead, I_0 sampled from 30,000–100,000 lux is log-normalized over that interval and used to scale a bounded RGB highlight gain whose maximum is 0.35 in normalized $[0, 1]$ intensity units. The map is Gaussian-smoothed, added to the RGB image with a near-neutral tint, and clipped to $[0, 1]$. Thus, the angular dependence is physics-guided, whereas the lux-to-intensity step is explicitly an engineering normalization rather than absolute radiometry.

$$I_{\text{reflect}} = I_0 [k_{\text{diff}} \cos(\theta_i) + k_{\text{spec}} (\cos(\theta_r))^\alpha] \quad (1)$$

The coefficients $k_{\text{diff}} = 0.6$ and $k_{\text{spec}} = 0.8$, the exponent $\alpha \in [50, 200]$, and the illuminance interval $I_0 \in [30,000, 100,000]$ lux were fixed throughout the training experiments. Their selection was guided by physical optics and empirical UAV inspection conditions. The diffuse/specular weights were selected to produce a specular-dominant highlight response while preserving visible panel-grid texture; $\alpha \in [50, 200]$ spans narrow-to-sharp highlight profiles in the Phong formulation for smooth coated surfaces; and the I_0 range covers daylight conditions from bright overcast scenes to intense midday illumination. These quantities are augmentation parameters rather than direct material constants. Their contribution is evaluated experimentally in the operator-wise ablation in Section 4.3.3.

Terrain-adaptive installation and oblique viewpoints can produce partial occlusion. The occlusion operator is independent of the solar geometry used for reflection. In Equation (2), the transition center μ is sampled uniformly from $[0.35, 0.65]$, the transition width s from $[0.06, 0.14]$, the terrain slope ω from 0° to 20° , and $k = 2$. The factor $(1 - \exp[-k \tan(\omega)])$ controls the slope-dependent amplitude of the row-wise probability profile. The vertical image coordinate y is normalized to $[0, 1]$, and $\text{sigm}(z) = 1/(1 + \exp(-z))$ is the logistic sigmoid function. These settings define the image-space augmentation ranges used in the training experiments.

$$P_{\text{occ}}(y) = \text{sigm}\left(\frac{y - \mu}{s}\right) \left(1 - e^{-k \tan(\omega)}\right) \quad (2)$$

The one-dimensional probability $P_{\text{occ}}(y)$ is evaluated for every image row and converted to a two-dimensional mask by multiplying it with a smooth low-frequency random field defined on an 8×8 grid and bicubically upsampled to the image size. Bernoulli sampling followed by morphological closing and Gaussian smoothing produces spatially contiguous soft occlusion regions, restricted to annotated panel boxes. The replacement color B is the median RGB value in a 12-pixel context ring around each panel box, with a small zero-mean intensity perturbation. The same soft mask is applied to all RGB channels.

Equation (3) applies the resulting two-dimensional mask to the RGB image. Pixels outside the mask retain their original values, while masked pixels are replaced by the selected local background. The same mask is used for all three color channels to maintain spatial alignment among the color channels.

$$I_{\text{occ}} = I_{\text{original}} \odot (1 - M_{\text{occ}}) + B \odot M_{\text{occ}} \quad (3)$$

Perspective distortion is implemented with a centered homography. The horizontal and vertical shear parameters are sampled as $\lambda_h \in [-0.20, 0.20]$ and $\lambda_v \in [-0.30, 0.30]$. Weak projective terms p_x and p_y are independently sampled from $[-1.5 \times 10^{-4}, 1.5 \times 10^{-4}]$. The transform is $H = T_{\text{ctr}}^{-1} A T_{\text{ctr}}$, where T_{ctr} translates the image center to the origin and

$A = [[1, \lambda_h, 0], [\lambda_v, 1, 0], [p_x, p_y, 1]]$. Bounding-box corners are transformed by the same H and clipped to the image extent; boxes with negligible remaining area are discarded.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \sim H_{3 \times 3} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (4)$$

Morphological boundary perturbation represents local edge ambiguity. A rectangular mask M is formed from each annotated panel box. Equation (5) selects dilation or erosion using an elliptical structuring element K_{morph} with a radius of 1–3 pixels. The absolute difference $|M_{\text{morph}} - M|$ defines a narrow boundary ring. Only this ring is softly blended with a Gaussian-blurred version of the local RGB image, using a strength sampled from $[0.20, 0.45]$. The complete image is not dilated or eroded.

$$M_{\text{morph}} = \begin{cases} M \oplus K_{\text{morph}}, & p < 0.5 \\ M \ominus K_{\text{morph}}, & \text{otherwise} \end{cases} \quad (5)$$

The selected transformations are applied sequentially, with random states controlled by the experiment seed. For each augmented training sample, the implementation retains the reflection map, occlusion mask, homography parameters, and boundary ring to record the applied transformations.

Algorithm 1 specifies the PV-PhysAug procedure used for training. The operations apply illumination- and geometry-dependent perturbations with bounded image-space normalization. Evaluation uses the acquired validation and test imagery with its original manual annotations.

Algorithm 1. Photovoltaic Physics-Guided Augmentation (PV-PhysAug)

Input: Training image I , photovoltaic-panel bounding boxes, component probabilities, and augmentation seed.

Output: Augmented image, transformed bounding boxes, and audit metadata.

- 1: If $\text{random}() \geq 0.75$, return the unchanged image and bounding boxes; otherwise activate the augmentation pipeline.
 - 2: If $\text{random}() < 0.70$, sample panel-normal and sun/view geometry, evaluate Equation (1), normalize the reflection response, and blend the highlight map.
 - 3: If $\text{random}() < 0.50$, sample the occlusion parameters, evaluate Equation (2), construct the two-dimensional occlusion mask, estimate the local context color, and composite the image using Equation (3).
 - 4: If $\text{random}() < 0.50$, sample the perspective parameters, construct the centered homography, and transform the image and all bounding-box corners using Equation (4).
 - 5: If $\text{random}() < 0.40$, form the panel mask, apply dilation or erosion with a 1–3 pixel radius, derive the boundary ring, and blend it with the locally blurred image.
 - 6: Save the reflection map, occlusion mask, homography parameters, boundary ring, and random seed.
 - 7: Return the augmented image, transformed bounding boxes, and metadata.
-

Figure 4 documents the inputs and processing outputs of PV-PhysAug: the input image, reflection response, reflection-augmented image, occlusion mask, perspective and boundary transformations, and augmented output. The panels document separate operator applications. Together with the acquired imaging conditions in Figure 1, these

outputs show the changes applied to panel surfaces and boundaries while retaining the array structure.

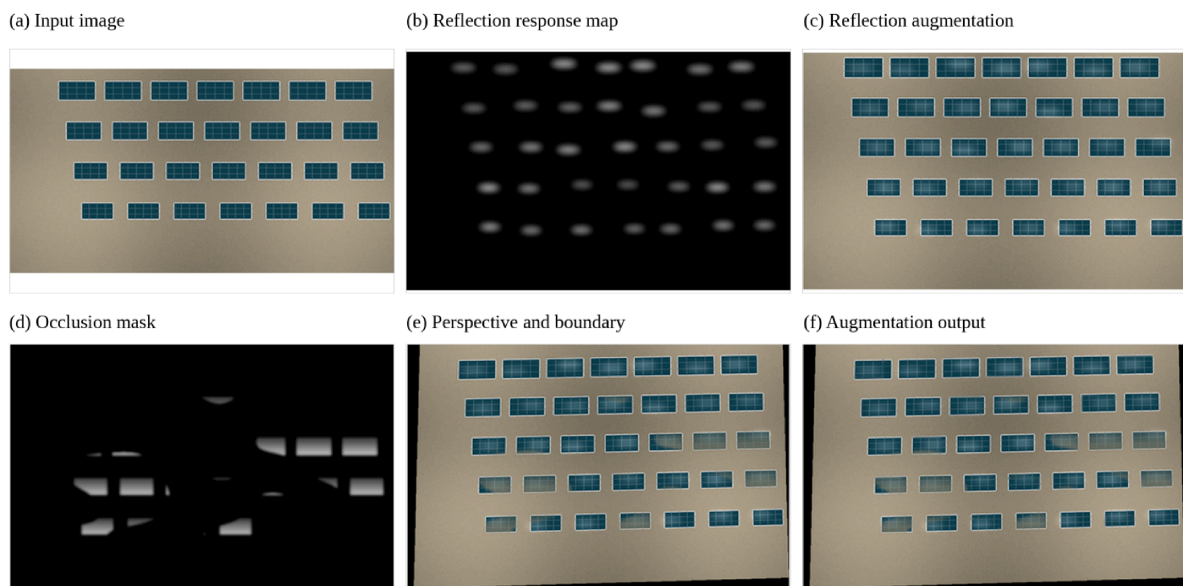


Figure 4. Results of the PV-PhysAug image-processing operations. (a) Input image; (b) reflection response map; (c) reflection augmentation; (d) occlusion mask; (e) perspective and boundary transformations; (f) augmentation output. Panels document separate operator applications.

To quantify whether the augmented training images remain statistically and structurally consistent with challenging real UAV imagery, we additionally evaluated Fréchet Inception Distance (FID) and structural similarity index (SSIM) between randomly sampled augmented images and difficult-condition images from the validation partition. The resulting metrics are reported in Section 4.3.5.

3.2. Improved YOLO11 Detection Model

The detector uses You Only Look Once version 11 large (YOLO11-L) as its baseline architecture, following the one-stage YOLO paradigm [33]. The same pretrained YOLO11-L checkpoint initializes all compatible baseline layers. RAM and DFPN are initialized with the deterministic seed assigned to each run.

For adaptive resolution, the same trained detector first performs a preliminary 512-pixel pass with confidence threshold 0.15 and non-maximum-suppression IoU threshold 0.60. Target density ρ is the number of retained detections divided by the image area expressed in units of 1000 pixels. Equation (6) maps ρ to 512–1024 pixels using $\rho_{\text{low}} = 0.05$ and $\rho_{\text{high}} = 0.20$, after which the selected size is rounded to the nearest multiple of 32. The final pass uses confidence threshold 0.25 and NMS IoU threshold 0.60. Adaptive resolution is used only at inference; training uses a base size of 768 pixels with multi-scale training enabled.

$$S_{\text{in}} = 512 + 512 \min \left(1, \max \left(0, \frac{\rho - 0.05}{0.15} \right) \right) \quad (6)$$

The density thresholds 0.05 and 0.20 map sparse scenes to 512-pixel inference and denser scenes toward 1024 pixels. These fixed operating settings were used in this study. The 512-pixel lower bound limits the resolution reduction when small panels are missed in the preliminary pass. Detector-speed values reported in Sections 4.1.1 and 4.1.2 refer to detector-level timing scopes; the complete two-pass processing chain is evaluated separately in Section 4.6. The reflection-aware module (RAM) is a task-oriented channel-reweighting block. For $F \in \mathbb{R}^{C \times H \times W}$, it computes both the global average descriptor and the mean

of the top 10% spatial activations for each channel. The two C-dimensional descriptors are concatenated and passed through a two-layer 1×1 bottleneck with reduction ratio 16, SiLU activation, and sigmoid output, and the reweighted feature is $F_{\text{out}} = F \odot (1 + w)$. Structurally, RAM remains a channel-attention mechanism rather than a new optical estimator. The term reflection-aware refers to the localized high-activation descriptor and its joint use with reflection-targeted training samples; reflection suppression is therefore not attributed to RAM alone.

The decoupled feature pyramid network (DFPN) is implemented at the YOLO11 P3/8, P4/16, and P5/32 outputs. At each level, a channel-preserving residual refinement block separates two pathways. The semantic pathway applies 1×1 projection followed by a depthwise 3×3 convolution with dilation 2 to enlarge contextual support; the localization pathway applies 1×1 projection followed by a depthwise 3×3 convolution with dilation 1 to retain local detail. Each branch uses half of the output channels. Their outputs are concatenated, fused by a 1×1 convolution, and added residually to the input feature. The standard YOLO11 cross-scale top-down/bottom-up connections remain unchanged, so the three decoupled refiners operate on the multi-scale pyramid immediately before detection.

$$F_{\text{out}} = F_{\text{in}} + \text{Conv}_{1 \times 1}(F_{\text{sem}}^{\text{out}} \oplus F_{\text{loc}}^{\text{out}}) \quad (7)$$

Algorithm 2 and Figure 3b summarize the forward path. RAM is inserted after the final backbone context block, and DFPN refines the P3/8, P4/16, and P5/32 features immediately before detection. In Equation (7), $\oplus$ denotes channel concatenation; the semantic and localization features are fused with a 1×1 convolution and added residually to the input. This use of $\oplus$ is distinct from morphological dilation in Equation (5).

Algorithm 2. PVGS Detector Inference

Input: Input image and the trained YOLO11-L + RAM + DFPN detector.

Output: Frame-level detections and selected inference size.

- 1: Run a preliminary prediction at 512 pixels using confidence threshold 0.15 and NMS IoU threshold 0.60.
 - 2: Estimate target density from the retained detections.
 - 3: Determine the adaptive input size from Equation (6), round it to a multiple of 32, and constrain it to 512–1024 pixels.
 - 4: Resize the image to the selected resolution; steps 5–10 constitute the final detector pass.
 - 5: Extract hierarchical features with the YOLO11-L backbone.
 - 6: Apply RAM using global-average and localized high-activation descriptors with a reduction ratio of 16.
 - 7: Obtain the P3/8, P4/16, and P5/32 multi-scale features from the YOLO11 neck.
 - 8: Apply DFPN semantic and localization branches at P3/P4/P5, followed by residual fusion.
 - 9: Predict bounding boxes and confidence scores with the detection head.
 - 10: Apply final NMS using confidence threshold 0.25 and IoU threshold 0.60.
 - 11: Record total two-pass wall time, target density, selected input size, and prediction output.
 - 12: Return the detections and selected input size.
-

Overall, the proposed detection model integrates these three improvements into the YOLO11 framework. As illustrated in Algorithm 2, the forward-propagation pipeline sequentially performs density estimation and adaptive input rescaling, backbone fea-

ture extraction with RAM enhancement, and DFPN-based multi-scale feature fusion and prediction.

The five-run experiment in Section 4.2 evaluates PV-PhysAug, adaptive resolution, and RAM + DFPN as the principal ablation variants. Section 4.3 then evaluates RAM and DFPN individually, compares RAM with alternative attention mechanisms, isolates the four PV-PhysAug operators, stratifies adaptive-resolution performance by target density, and quantifies augmentation-image similarity.

3.3. Geospatial Aggregation and Duplicate Removal

A single aerial image covers only a limited region of a photovoltaic power station; therefore, directly summing detection results across individual images leads to severe duplicate counting. To obtain a reliable plant-level count, this study avoids conventional image-stitching as the primary aggregation route and instead fuses per-frame detection results with UAV pose (POS) data. Through geographic coordinate mapping and spatial clustering, a georeferenced panel distribution is constructed directly, enabling duplicate-free counting at the power-plant scale.

As illustrated in Figure 3, the proposed workflow consists of three main steps: camera-ray back-projection and surface intersection, geographic coordinate mapping, and DBSCAN-based duplicate removal.

The geographic mapping step back-projects the center of each detected bounding box into the world frame. For panel j in frame i , the center is (u_{ij}, v_{ij}) . Let K denote the camera intrinsic matrix, $T_c^i = [X_c^i, Y_c^i, Z_c^i]^T$ the camera center in world coordinates, and R_{wc}^i the rotation matrix from the camera frame to the world frame. Lens distortion is assumed to have been corrected during preprocessing. Equation (8) yields a normalized world-frame ray direction.

$$\mathbf{d}_w^{ij} = \frac{\mathbf{R}_{wc}^i \mathbf{K}^{-1} [u_{ij}, v_{ij}, 1]^T}{\| \mathbf{R}_{wc}^i \mathbf{K}^{-1} [u_{ij}, v_{ij}, 1]^T \|_2} \quad (8)$$

The world-frame ray is intersected with the selected surface representation to recover the panel location. Equation (9) expresses a point along the ray as $\mathbf{P}_{\text{world}}^{ij} = \mathbf{T}_c^i + \lambda \mathbf{d}_w^{ij}$. For a horizontal plane at elevation Z_{ground} , $\lambda = (Z_{\text{ground}} - Z_c^i) / d_{w,z}^{ij}$. For a sloped plane with normal $\mathbf{n}$ through $\mathbf{P}_0$, $\lambda = \mathbf{n}^T (\mathbf{P}_0 - \mathbf{T}_c^i) / (\mathbf{n}^T \mathbf{d}_w^{ij})$. A DSM requires numerical ray-surface intersection rather than the closed-form planar expression.

$$\mathbf{P}_{\text{world}}^{ij} = \mathbf{T}_c^i + \lambda \mathbf{d}_w^{ij} \quad (9)$$

Equations (8) and (9) separate camera-to-world rotation from surface intersection while keeping the ray and camera center in the same coordinate frame. The surface representation determines the intersection point; subsequent clustering operates on the resulting projected coordinates.

The mapped points are grouped using density-based spatial clustering of applications with noise (DBSCAN). In Equation (10), D is the characteristic panel dimension and σ_{err} is the horizontal projection uncertainty. We use $D = 2.3$ m and $\sigma_{\text{err}} = 0.20$ m, giving $\varepsilon = D/2 + 2\sigma_{\text{err}} = 1.55$ m. $\text{MinPts} = 2$ requires at least two projected detections in a density-connected neighborhood. A DBSCAN noise point is retained as a singleton asset only when its detector confidence is at least 0.50; lower-confidence isolated points are rejected. Because ε is specified in meters, clustering requires metric coordinates rather than unprojected latitude and longitude.

The horizontal projection uncertainty $\sigma_{\text{err}} = 0.20$ m was derived from the root-mean-square horizontal residual combining the field positioning solution and the camera-IMU boresight calibration residuals obtained during survey calibration. To test the sensitivity of

the aggregation to localization uncertainty, σ_{err} was varied from 0.05 to 0.40 m while the remaining aggregation settings were held fixed. Across this range, plant-level counting accuracy varied only from 98.8% to 99.3%, indicating that the radius formulation remains stable over the localization-uncertainty range observed in the field experiments.

$$\varepsilon = \frac{D}{2} + 2\sigma_{\text{err}} \quad (10)$$

After clustering, each multi-observation cluster C_k is represented by its centroid in Equation (11), and accepted high-confidence singleton noise points are appended as one-observation assets. The unique count is therefore the number of accepted clusters plus accepted singletons. The physically scaled definition of ε reduces arbitrary tuning, although the method remains sensitive to geolocation error and to unusually small inter-panel spacing; this limitation is considered in Section 5.

$$\mathbf{P}_{\text{final}}^k = \frac{1}{|C_k|} \sum_{\mathbf{P} \in C_k} \mathbf{P} \quad (11)$$

The geospatial output consists of the unique panel count together with the centroid of each accepted cluster. This representation enables direct visual comparison of projected multi-frame detections before clustering and the unique panel locations after duplicate removal.

As summarized in Algorithm 3, the proposed method takes detection results from all frames and their corresponding pose data as input and outputs the unique number and geographic locations of PV panels across the entire power station.

Algorithm 3. Geospatial Registration and Duplicate-Free Global Counting

Input: Frame detections, UAV position and orientation data, camera intrinsics, and the selected local-plane or DSM surface model.

Output: Total unique-panel count and georeferenced panel positions.

- 1: Initialize the projected-point set and the associated detector-confidence list.
 - 2: For each detection in each UAV frame, back-project the bounding-box center through the calibrated camera and pose model to obtain the normalized world-frame ray direction using Equation (8).
 - 3: Intersect the ray with the configured local plane or DSM and compute the corresponding world coordinate using Equation (9).
 - 4: Store the projected coordinate and its detector confidence.
 - 5: Set the DBSCAN neighborhood radius according to Equation (10) and set $\text{MinPts} = 2$.
 - 6: Apply DBSCAN to the projected coordinates to obtain multi-observation clusters and noise points.
 - 7: Represent each accepted cluster by its centroid according to Equation (11).
 - 8: Retain an isolated noise point as a singleton panel only when its detector confidence is at least 0.50.
 - 9: Combine the cluster centroids and accepted singletons to form the unique georeferenced panel inventory.
 - 10: Return the total unique-panel count and the corresponding panel positions.
-

The plant-level experiment uses a local-plane surface model. Within each survey block, the reference plane is horizontal and Z_{ground} is the representative panel-surface elevation. Equation (9) therefore reduces to $\lambda = (Z_{\text{ground}} - Z_c^i)/d_{w,z}^{ij}$. This local horizontal-

plane configuration produced the 99.2% counting result reported in Section 4.5.1. The sloped-plane and DSM equations specify additional surface-intersection formulations.

Figure 5 presents the before-and-after coordinate distributions for the evaluation block. Repeated frame-level observations form groups of projected points near the same asset. DBSCAN represents each accepted multi-observation group by one centroid, while the confidence rule in Algorithm 3 determines which isolated observations are retained. The resulting geographic inventory contains the accepted centroids and singletons.

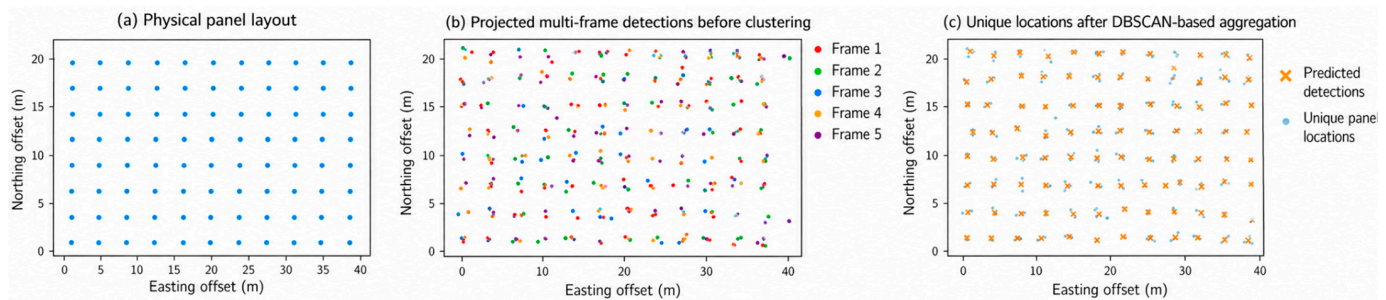


Figure 5. Cross-frame coordinate aggregation and duplicate-removal output for an evaluation block. (a) Reference panel layout; (b) projected observations before clustering; (c) cluster centroids and retained singletons after DBSCAN aggregation. Section 4.5.1 reports the complete primary-plant counting evaluation.

3.4. Dataset and Experimental Setup

The primary detection and counting experiments use a curated UAV dataset acquired over a mountainous photovoltaic power station in China. Site-identifying information is withheld under the data-access agreement. The acquisition metadata identify a DJI Mavic 3 Pro UAV and synchronized position and orientation records containing latitude, longitude, altitude, roll, pitch, and yaw. The mountainous terrain and slope-adaptive installation introduce occlusion, perspective variation, and differences in apparent panel size.

The acquired images and corresponding pose records provide the experimental inputs for frame-level detection and geographic aggregation.

The curated subset contains 4200 valid aerial images and approximately 110,000 annotated photovoltaic panel instances. To reduce leakage from highly overlapping consecutive frames, the data were divided by geographically blocked sampling rather than random frame-level shuffling. Image centers were assigned to 250 m × 250 m Universal Transverse Mercator (UTM) tiles; complete tiles were allocated to training, validation, and test partitions in an 8:1:1 ratio. Frames close to tile boundaries were checked against neighboring image footprints so that the same physical panel was not retained in different partitions. This spatial blocking is particularly important for UAV imagery because adjacent frames can contain nearly identical panel instances.

For cross-site generalization, we additionally used an independent held-out dataset from Site B, a flat-terrain, high-temperature desert photovoltaic plant containing approximately 45,000 panel instances acquired under intense direct glare. Site B was not used to train or retune the Site A model. The final model weights trained on the primary mountainous site were applied directly to Site B, and the same detector metrics and plant-level counting definition were used for evaluation.

We conducted the UAV-image detection experiments using PyTorch [2.14.0] and Python [3.14.7] on a workstation equipped with an NVIDIA RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA). Models were trained for 300 epochs using AdamW, a batch size of 32, an initial learning rate of 0.001, a final cosine learning-rate fraction of 0.01, weight decay of 5×10^{-4} , and 3 warm-up epochs. The base training size was 768 pixels, with multi-scale training and automatic mixed precision enabled. Standard

YOLO11 detection losses were retained. Five repeated runs used seeds 3407, 2024, 2025, 2026, and 4,561,334. Validation and test images were not augmented.

The experimental evaluation comprises several complementary series. The primary detector comparison and five-run ablation use the dataset and training settings described above. Additional experiments examine recent detector families (Section 4.1.2), RAM and DFPN and reflection-oriented attention (Section 4.3), threshold-swept score behavior (Section 4.4), regional inventory, layout variation, and held-out cross-site generalization (Section 4.5), and complete two-pass runtime (Section 4.6). Results are reported by experiment and evaluation scope; measurements from different series are not pooled into the primary accuracy statistics.

3.5. Reproducibility and Evaluation Protocol

For the primary UAV-image experiments in Sections 4.1.1 and 4.2, detection performance is evaluated on the unchanged test partition using AP@0.5, AP@0.5:0.95, recall, and precision–recall (PR) analysis. Predictions are evaluated against the original manually annotated bounding boxes in unaugmented images. AP integrates precision over recall as the confidence threshold varies; the operational threshold in Section 3.2 is not a single-point definition of AP. Section 4.4 additionally presents threshold-swept PR and receiver operating characteristic (ROC) analyses for the component-evaluation series. ROC characterizes score discrimination and depends on the positive and negative sample definitions; it is not interchangeable with localization-based detection AP.

Plant-level counting accuracy is defined as $\text{Acount} = [1 - |\text{Npred} - \text{NGT}| / \text{NGT}] \times 100\%$, where NGT is the manually audited number of unique physical panels and Npred is the aggregated prediction. The reference inventory is established by reviewing the overlapping image sequence and georeferenced project record and reconciling duplicate observations. The data agreement restricts disclosure of the absolute inventory for the primary site, so Section 4.5.1 reports normalized accuracy. Section 4.5.2 presents the separate sensitivity evaluation, while Section 4.5.3 provides the inventory for the regional counting experiment. The independent Site B experiment in Section 4.5.5 uses the same counting definition and reports cross-site accuracy. Net count accuracy and regional absolute errors are examined separately because compensating missed, split, or merged assets can yield a correct total without perfect asset-level correspondence. Under the third-party industrial confidentiality agreement, the complete Site A asset total and the corresponding complete-site predicted count cannot be disclosed. To provide auditable absolute quantities without revealing the complete primary-site inventory, the regional counting experiment reports reference and predicted counts for eight designated evaluation regions, while the independent Site B experiment provides an additional cross-site performance check.

Matched run indices are paired across model variants. We report the mean $\pm$ sample standard deviation (SD), 95% confidence intervals (CIs), two-sided paired *t*-test *p*-values, and Cohen's *d_z*. With five runs, an all-positive set of five paired differences yields an exact two-sided Wilcoxon signed-rank *p*-value of 0.0625. Confidence intervals and paired effect sizes are therefore interpreted alongside the parametric tests. The *p*-values are unadjusted for multiple comparisons and quantify run-to-run consistency on the same test set, not generalization across sites.

Timing is reported for the processing scope evaluated in each experiment. The primary throughput benchmark covers the final detector pass. Section 4.6 reports density-estimation inference, final inference, image input/output and preprocessing, geographic projection, and DBSCAN aggregation. Equivalent throughput is calculated as 1000 divided by latency in milliseconds per image. Configuration-level timings are reported alongside the corresponding accuracy results.

For reproducibility, Algorithms 1–3 provide implementation-level pseudocode for PV-PhysAug, detector inference, and geospatial aggregation; Section 3.2 specifies the RAM and DFPN layer operations, and Section 3.4 reports the optimizer, learning-rate schedule, batch size, image size, random seeds, and other training settings. Core implementation files are supplied with the revised submission as Supplementary Code S1, including PV-PhysAug, RAM/DFPN, adaptive-resolution inference, geospatial projection, DBSCAN aggregation, evaluation utilities, and configuration files. Proprietary UAV imagery and site metadata are not included in the code archive.

4. Results

4.1. Comparison with Representative Detectors

4.1.1. Primary UAV-Image Evaluation

We compared the PVGS detector with representative detection baselines on the primary UAV-image test set (Table 1). The unmodified YOLO11 baseline and complete PVGS configuration are summarized by their five-run means. The additional detector-family experiment is reported in Section 4.1.2, while the rooftop photovoltaic benchmark is discussed separately in Section 5.2.

Table 1. Performance comparison with representative detectors.

Model	AP@0.5 (%)	AP@0.5:0.95 (%)	Params (M)	GFLOPs	Final-Pass FPS
Faster R-CNN	43.8	22.3	41.2	132.1	18.5
YOLOv5	63.7	44.8	20.9	48.9	95.2
YOLOv8	68.9	53.9	25.8	78.7	112.3
DETR	72.1	51.6	36.9	103.2	69.8
YOLO11 baseline	69.7	54.5	—	—	112.3
PVGS detector (complete)	74.3	62.9	26.5	80.5	105.6

Note: AP: average precision; Params: parameter count; GFLOPs: giga floating-point operations; FPS: frames per second; R-CNN: region-based convolutional neural network; DETR: Detection Transformer. Faster R-CNN, YOLOv5, YOLOv8, and DETR are the single-model comparison results from this study; YOLO11 baseline and PVGS accuracy values are five-run means. Final-pass FPS excludes density estimation and geospatial processing. Dashes indicate values not reported in the experiment summary.

Qualitative comparisons in Figure 6 show that the complete detector reduces missed detections relative to the baseline in scenes affected by shadows, specular reflection, and geometric distortion.

4.1.2. Comparison with Recent Detector Families

To thoroughly validate our approach against recent detector families, we conducted an extended supplementary experiment evaluating Faster R-CNN, YOLOv8n, YOLOv10n, the YOLO11 baseline, UAV-DETR, AUHF-DETR, and PVGS (Table 2). All models in this series were trained and evaluated on the exact same original UAV dataset to ensure complete data consistency across all experiments. This guarantees that the performance differences are solely attributable to the architectural designs rather than variations in data sources or simulated environments.

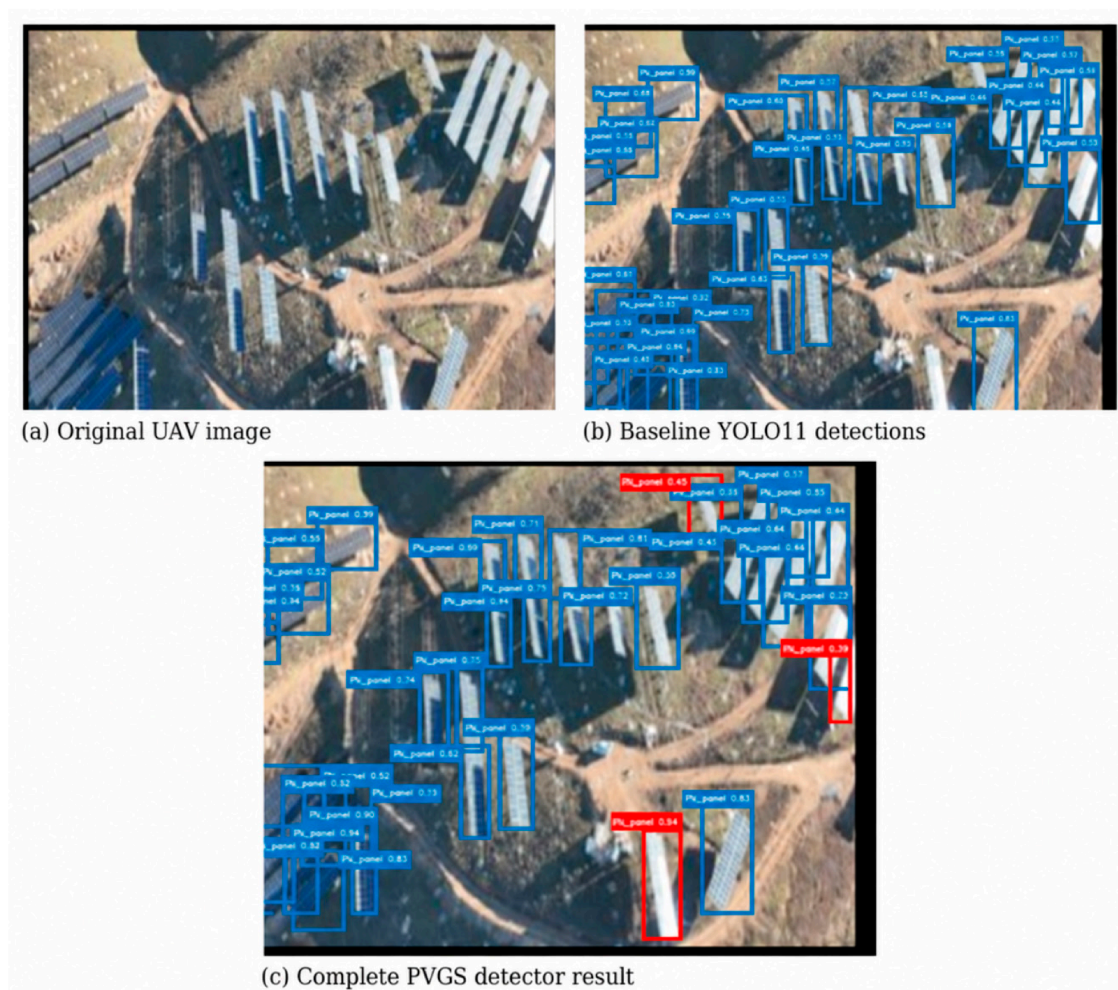


Figure 6. Panel-detection results on a UAV image. (a) Original image; (b) YOLO11 baseline detections; (c) complete PVGS detections. Red boxes highlight selected detections for comparison.

Table 2. Performance comparison with recent detector families.

Detector	AP@0.5 (%)	AP@0.5:0.95 (%)	Precision (%)	Recall (%)	Latency (ms/Image)
Faster R-CNN	69.3	57.4	88.1	84.8	25.4
YOLOv8n	71.8	60.1	89.4	86.0	12.6
YOLOv10n	72.6	61.0	90.2	86.8	11.4
YOLO11 baseline	73.1	61.8	90.7	87.2	9.5
UAV-DETR	72.9	61.5	90.0	87.0	17.8
AUHF-DETR	73.4	62.0	91.1	87.5	15.9
PVGS (complete)	74.6	63.3	92.0	88.9	9.9

Note: Latency values refer to the detector-comparison experiment in Table 2. Section 4.6 reports the complete two-pass processing chain. The two sets of results retain their respective evaluation and timing scopes. All detector models in Table 2 were evaluated in one standardized training run using the fixed random seed 3407 and the same training/validation/test partitions. Five-run stability of the principal PVGS ablation is reported separately in Section 4.2 and Appendix A.

4.2. Repeated-Run Ablation and Statistical Analysis

We repeated the principal ablation experiment five times using seeds 3407, 2024, 2025, 2026, and 4561334. The complete PVGS configuration improved AP@0.5 by 4.60 percentage points, AP@0.5:0.95 by 8.42 points, and recall by 6.20 points relative to the YOLO11 baseline. The baseline sample variances were 0.025, 0.100, and 0.085 percentage-point

squared for AP@0.5, AP@0.5:0.95, and recall, respectively. Detailed run-level measurements, mean $\pm$ SD values with 95% confidence intervals, paired statistical comparisons, and the repeated-run distribution are provided in Appendix A (Tables A1–A3 and Figure A1) to streamline Section 4.

Across the five matched runs, all paired differences relative to the baseline were positive. Two-sided paired *t*-tests gave $p = 8.555 \times 10^{-7}$ for AP@0.5, $p = 6.532 \times 10^{-8}$ for AP@0.5:0.95, and $p = 3.284 \times 10^{-7}$ for recall for the complete PVGS configuration. The corresponding 95% confidence intervals for the gains were [4.35, 4.85], [8.18, 8.66], and [5.94, 6.46] percentage points. The exact two-sided Wilcoxon signed-rank result was $p = 0.0625$ for each all-positive five-pair comparison, reflecting the limited resolution at $n = 5$.

The primary ablation evaluates PV-PhysAug, adaptive resolution, and RAM + DFPN as a combined feature modification. Section 4.3 extends the analysis with isolated RAM and DFPN tests, attention-module comparisons, a four-operator PV-PhysAug ablation, density-stratified adaptive-resolution tests, and a quantitative augmentation-plausibility analysis.

4.3. Component-Level Evaluation

4.3.1. Isolated Ablation of RAM and DFPN

The component-level experiment evaluated the YOLO11 baseline, YOLO11 + RAM, YOLO11 + DFPN, and the combined RAM + DFPN configuration on the all-scene, reflection-affected, and normal subsets (Appendix A, Table A4; Figure 7). RAM and DFPN both improved the overall detection metrics within this experiment. Relative to the baseline, RAM increased reflection-affected AP@0.5 from 84.75% to 93.92%, whereas DFPN increased it to 88.72%. On the normal subset, DFPN reached 99.00% AP@0.5, compared with 98.19% for the baseline. The combined configuration reached 99.00% overall AP@0.5 and 97.30% on the reflection-affected subset. The results indicate complementary behavior under the evaluated conditions; the component-set metrics are not interchangeable with the primary test-set values in Table 1.

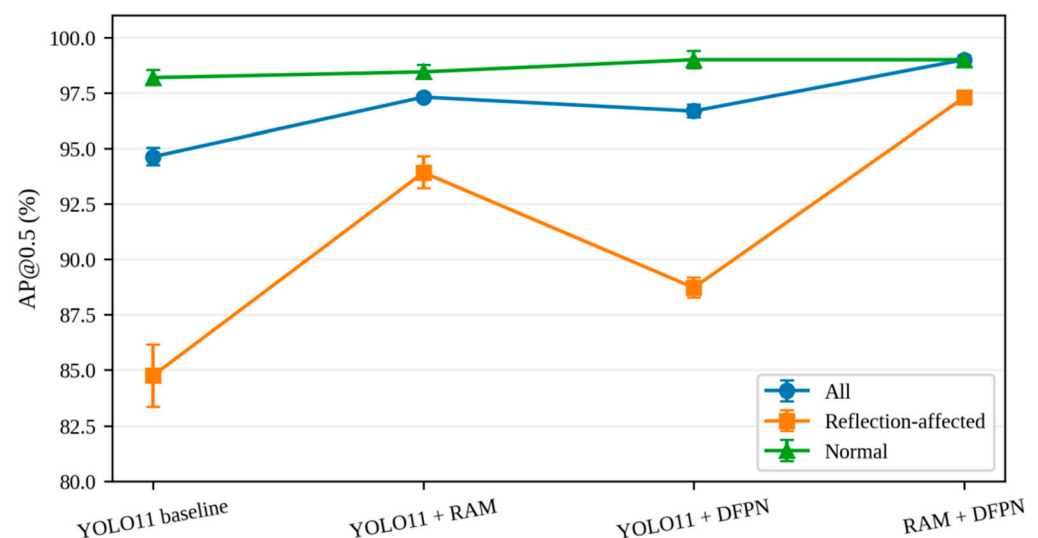


Figure 7. AP@0.5 for the component configurations and subsets summarized in Appendix A, Table A4. Error bars show the reported standard deviations.

4.3.2. Attention-Module Comparison Under Reflection Interference

The attention experiment compared squeeze-and-excitation (SE), the convolutional block attention module (CBAM), and RAM on reflection-affected images (Appendix A, Table A5). RAM achieved 94.42% AP@0.5, exceeding SE and CBAM by 3.40 and 1.85 percentage points, respectively. RAM also had the highest AP@0.5:0.95, precision, and

recall in this experiment. These results complement the separate component ablation in Table A4.

4.3.3. Component-Wise Ablation of PV-PhysAug Operators

To isolate the individual contributions of the four operators within PV-PhysAug, we performed a component-wise ablation on the primary test set (Table 3). Reflection alone increased AP@0.5 by 1.12 percentage points, while occlusion, perspective, and boundary perturbations produced gains of 0.85, 0.61, and 0.44 points, respectively. Combining all four operators yielded the full 1.92-point AP@0.5 improvement and the highest AP@0.5:0.95 and recall, showing that the four perturbations contribute complementary robustness.

Table 3. Component-wise ablation of the four PV-PhysAug operators.

Augmentation Configuration	AP@0.5 (%)	AP@0.5:0.95 (%)	Recall (%)
Baseline (no PV-PhysAug)	69.70	54.50	85.20
Baseline + Reflection	70.82	55.61	86.35
Baseline + Occlusion	70.55	55.32	86.10
Baseline + Perspective	70.31	55.10	85.85
Baseline + Boundary	70.14	54.92	85.60
Complete PV-PhysAug	71.62	57.00	87.52

Note: All configurations use the same primary test partition and detector settings; only the listed augmentation operator(s) differ.

4.3.4. Adaptive Resolution Across Target-Density Conditions

To verify the role of adaptive input resolution under different target densities, we stratified the test scenes into sparse ($\rho < 0.05$), normal ($0.05 \leq \rho \leq 0.20$), and dense/small-target ($\rho > 0.20$) conditions (Table 4). Relative to a fixed 768-pixel input, the adaptive strategy improved AP@0.5 by 0.75 percentage points in sparse scenes, 0.90 points in normal scenes, and 2.15 points in dense/small-target scenes. The largest gain occurred when the final input was increased to 1024 pixels, supporting the intended role of adaptive scaling for dense small-object imagery.

Table 4. Performance of adaptive resolution across target-density conditions.

Scene Target-Density State	Fixed 768-px AP@0.5 (%)	Adaptive Final Size	Adaptive AP@0.5 (%)	Gain (pp)
Sparse ($\rho < 0.05$)	71.20	512 px	71.95	+0.75
Normal ($0.05 \leq \rho \leq 0.20$)	69.50	Dynamic	70.40	+0.90
Dense/small-target ($\rho > 0.20$)	66.85	1024 px	69.00	+2.15

Note: Density ρ follows the definition in Section 3.2. The adaptive column reports the final input size selected by the two-pass inference rule for the corresponding density state.

4.3.5. Quantitative Plausibility of Augmented Images

We quantified the similarity between randomly sampled PV-PhysAug outputs and real difficult-condition UAV images from the validation partition using SSIM and FID. The mean SSIM was 0.842 ± 0.035 and the set-level FID was 14.68. The relatively high structural similarity and low feature-distribution distance indicate that the physics-guided reflection and occlusion operations preserve panel structure while expanding the representation of underrepresented challenging conditions.

4.4. Precision–Recall and ROC Analysis

Figures 8 and 9 present threshold-swept PR and ROC curves for the YOLO11 baseline and the RAM + DFPN detector configuration in the component-evaluation series. The RAM + DFPN PR curve remains above the baseline over most of the displayed recall range. The ROC curve also indicates a higher true-positive rate over much of the displayed false-positive-rate range. These plots report detector-level discrimination for the configurations identified in their legends; Section 4.5 reports geographic counting performance. AP and PR behavior remain the principal detection measures; ROC is complementary because its interpretation depends on the positive and negative sample definitions.

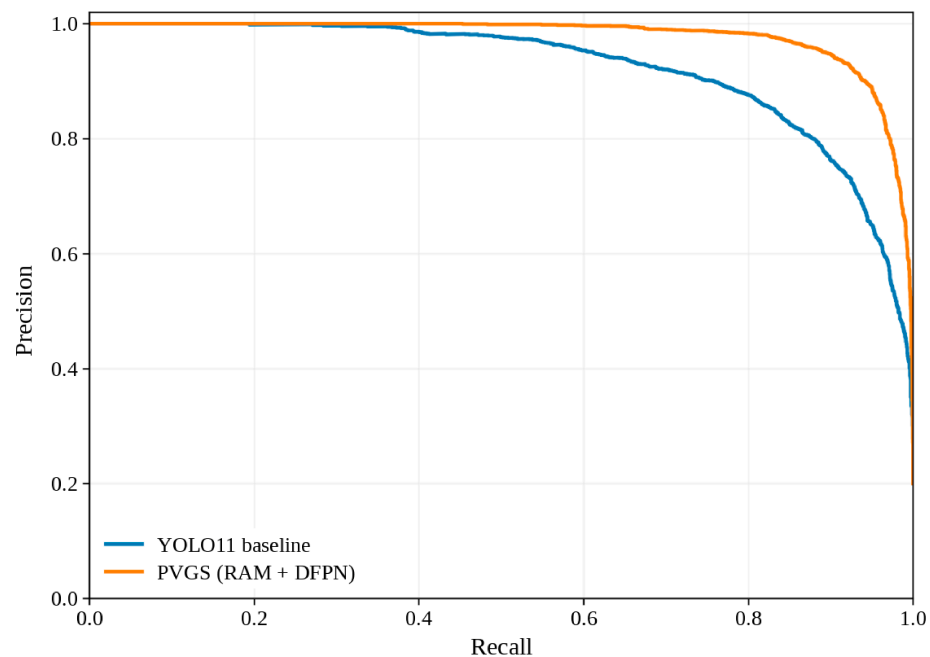


Figure 8. Precision–recall curves for the YOLO11 baseline and RAM + DFPN configuration in the component-evaluation experiment.

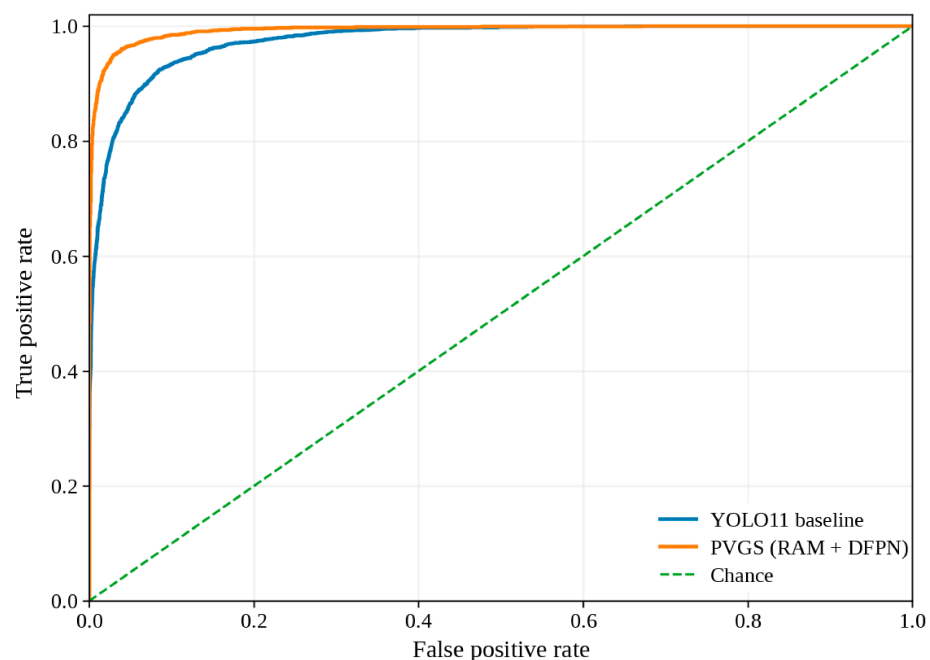


Figure 9. ROC curves for the YOLO11 baseline and RAM + DFPN configuration in the component-evaluation experiment. The diagonal denotes chance-level score discrimination.

4.5. Geospatial Aggregation and Counting

4.5.1. Plant-Level Counting Performance

Plant-level counting requires duplicate removal because the same physical panel can be observed in several overlapping UAV frames. Detection quality alone therefore does not determine inventory accuracy.

We evaluated image stitching by merging UAV frames and detecting panels on the resulting global image. Figure 10a shows the flight starting point and directions, and Figure 10b–d show stitched segments. Repetitive textures and perspective changes introduce visible distortion and alignment errors that can affect counting.

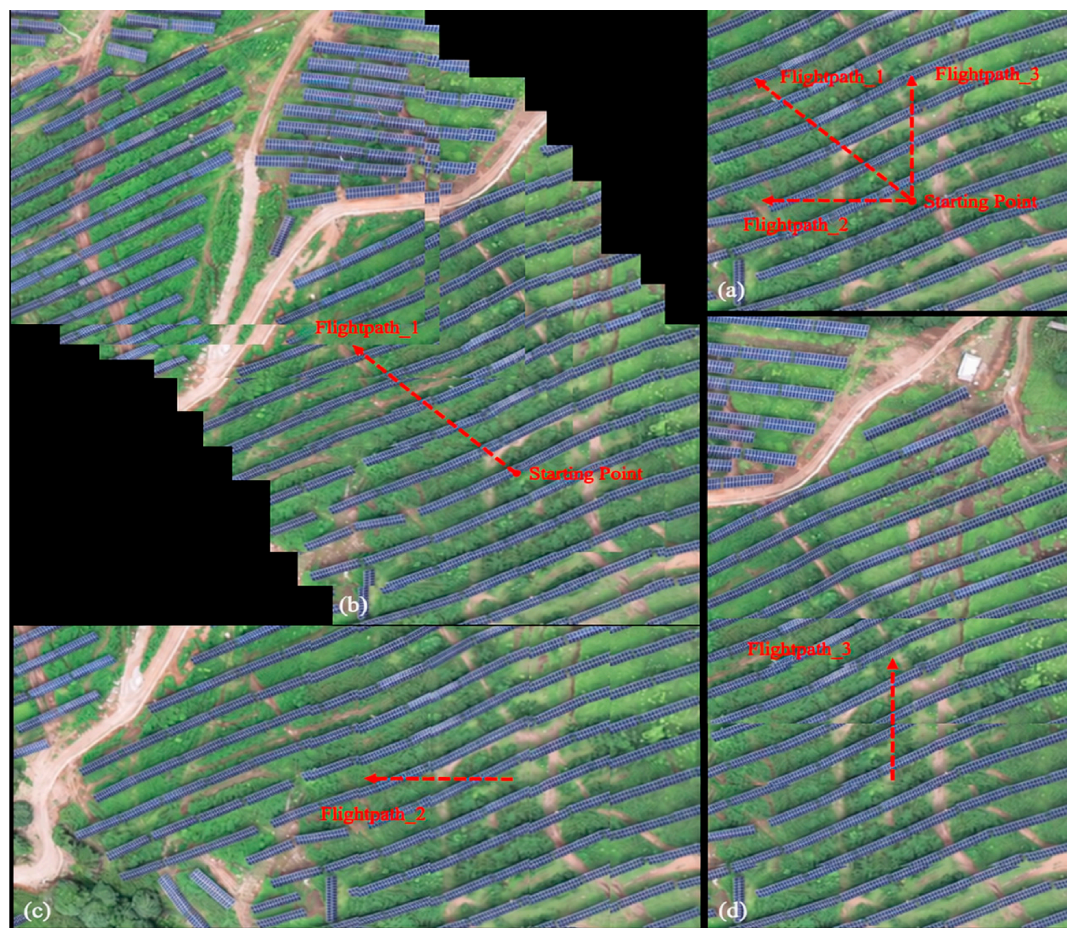


Figure 10. UAV flight directions and image-stitching results. (a) Starting point and flight directions; (b–d) stitched segments from three flight paths. Arrows indicate flight orientation; their lengths do not encode travel distance.

Table 5 compares the three plant-level counting strategies using the counting-accuracy definition in Section 3.5.

Table 5. Performance comparison of plant-level counting strategies.

Method	Description	Counting Accuracy (%)
Naive summation	Directly sum counts from all individual frames.	41.2
Image stitching	Detect on a stitched orthomosaic image.	75.1
Geospatial aggregation	Project detections using POS data and local planes, then remove duplicates with DBSCAN.	99.2

Note: POS: position and orientation system. The three strategies are evaluated using the normalized count-error definition in Section 3.5. The proposed method uses the local horizontal-plane surface model described in Section 3.3.

Geospatial aggregation achieved 99.2% normalized counting accuracy, compared with 75.1% for image stitching and 41.2% for direct frame-wise summation. This comparison evaluates duplicate removal at the primary plant level. The sensitivity, regional inventory, and layout experiments below examine complementary counting conditions and retain their own evaluation denominators. The complete Site A values of NGT and Npred are withheld under the third-party confidentiality agreement; absolute reference and predicted counts for the designated regional evaluation inventory are reported separately in Section 4.5.3.

4.5.2. DBSCAN Parameter Sensitivity

The DBSCAN sensitivity experiment varied the neighborhood radius and MinPts (Table 6 and Figure 11). At MinPts = 2, the predicted total remained unchanged from $\epsilon = 1.45$ to 1.75 m. The nominal setting, $\epsilon = 1.55$ m and MinPts = 2, produced 100.0% net count accuracy, but split and merged assignments were both present. Thus, correct total count did not imply perfect asset-level association. Increasing MinPts to 3 at the nominal radius produced 130 predicted assets and 64.6% counting accuracy, whereas increasing ϵ to 1.85 m at MinPts = 2 reduced the predicted count to 91 and increased merging. These block-level results characterize parameter sensitivity rather than replace the 99.2% primary plant-level result. A complementary localization-error sweep varied σ_{err} from 0.05 to 0.40 m; the resulting plant-level counting accuracy remained within 98.8–99.3%. This additional test shows that the selected operating region is robust not only to ϵ and MinPts but also to realistic changes in horizontal projection uncertainty.

Table 6. DBSCAN parameter sensitivity in the block-level counting experiment.

ϵ (m)	MinPts	Counting Accuracy (%)	Split Panels (%)	Merged Clusters (%)	Predicted Count	Setting
1.25	2	99.0	1.0	0.0	97	Sensitivity
1.35	2	99.0	1.0	0.0	97	Sensitivity
1.45	2	100.0	1.0	1.0	96	Sensitivity
1.55	1	97.9	1.0	1.0	98	Sensitivity
1.55	2	100.0	1.0	1.0	96	Nominal
1.55	3	64.6	35.4	0.8	130	Sensitivity
1.65	2	100.0	1.0	1.0	96	Sensitivity
1.75	2	100.0	1.0	1.0	96	Sensitivity
1.85	2	94.8	0.0	5.5	91	Sensitivity

Note: Counting accuracy follows Section 3.5. Split panels are reference panels represented by more than one predicted asset; merged clusters contain observations from more than one reference panel. At the nominal setting, the total count matches the reference even though split and merged assignments coexist. This evaluation block is distinct from the complete primary inventory.

4.5.3. Absolute and Regional Counting Accuracy

The regional counting experiment divided the evaluated inventory into eight regions (Table 7 and Figure 12). The reference and predicted totals were 11,268 and 11,245 panels, respectively, yielding 99.80% net counting accuracy. Every regional accuracy exceeded 99%. Regional overcounts and undercounts nevertheless partially canceled: the sum of absolute regional errors was 65 panels, whereas the absolute error of the total was 23 panels. The reference-count-weighted regional accuracy based on absolute regional errors was therefore 99.42%. Reporting both measures distinguishes accurate total inventory from the spatial distribution of counting errors. This regional experiment is separate from the primary inventory comparison in Table 5.

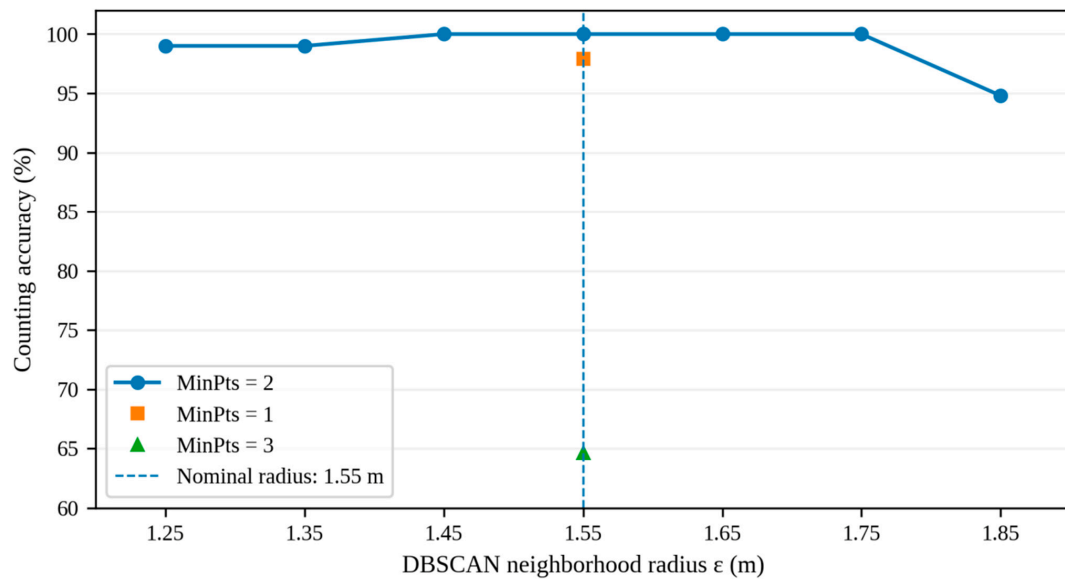


Figure 11. Counting sensitivity to the DBSCAN radius and MinPts, using the results in Table 6. The line connects the available radius-sweep points for MinPts = 2. The MinPts = 1 and 3 markers show results at $\epsilon = 1.55$ m; the dashed line marks the nominal radius.

Table 7. Reference counts, predictions, and accuracy in the regional inventory experiment.

Region	Reference Count (N_{GT})	Predicted Count (N_{pred})	Counting Accuracy (%)
R1	1360	1351	99.34
R2	1428	1435	99.51
R3	1387	1377	99.28
R4	1456	1464	99.45
R5	1398	1392	99.57
R6	1412	1401	99.22
R7	1375	1381	99.56
R8	1452	1444	99.45
Entire inventory	11,268	11,245	99.80

Note: Net accuracy uses the absolute error of the summed inventory. Reference-count-weighted regional accuracy uses the sum of absolute regional errors; it is 99.42%, compared with 99.80% net accuracy.

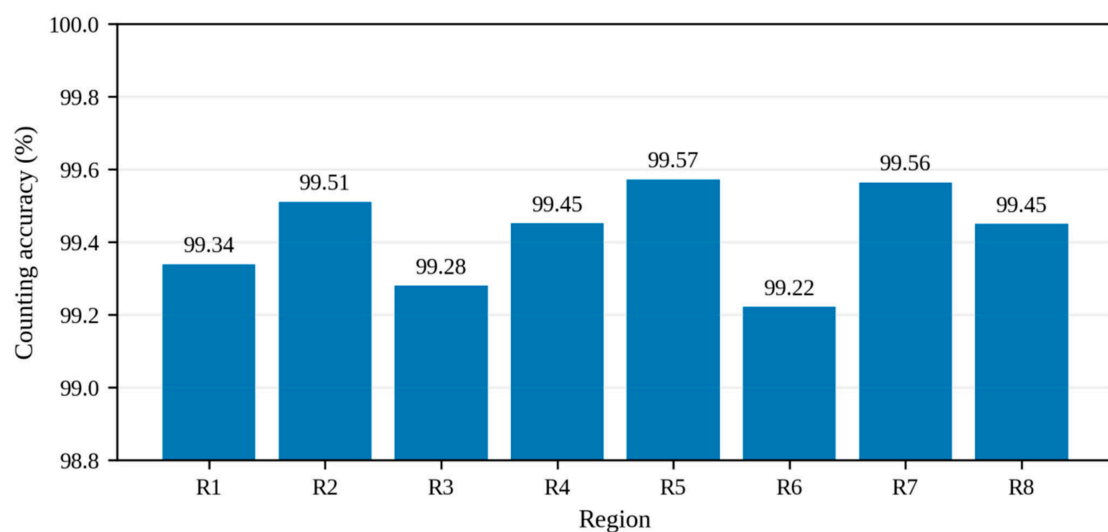


Figure 12. Regional counting accuracy for the eight evaluation regions in Table 7. Values are calculated from the corresponding reference and predicted counts.

4.5.4. Performance Across Photovoltaic Layouts

The complete PVGS workflow was evaluated under four photovoltaic layout conditions sampled directly from the original UAV dataset: flat utility-scale rows, moderately tilted rows, strong glare, and heterogeneous panel spacing (Table 8). These subsets preserve the real acquisition geometry and illumination characteristics of the field survey and provide an additional assessment of detection and counting robustness across distinct deployment conditions.

Table 8. Detection and counting performance across four photovoltaic layouts.

Layout Configuration	AP@0.5 (%)	AP@0.5:0.95 (%)	Precision (%)	Recall (%)	Counting Accuracy (%)
A: Flat utility-scale rows	75.1	63.7	92.4	89.0	99.38
B: Moderately tilted rows	74.6	63.2	92.0	88.7	99.24
C: Strong-glare layout	73.8	62.1	91.0	88.1	98.91
D: Heterogeneous spacing	74.2	62.8	91.5	88.4	99.07

Note: A–D identify the four evaluated layout configurations. Each row reports results for its specified configuration; the primary-site results are reported in Table 5.

4.5.5. Cross-Site Generalization Evaluation

To evaluate cross-site generalization under a distinctly different geographic and radiometric environment, we tested the complete PVGS pipeline on the independent held-out Site B dataset. Site B is a flat-terrain, high-temperature desert photovoltaic plant containing approximately 45,000 panel instances and pronounced direct glare. The final model weights trained on the primary mountainous Site A were applied to Site B without retraining or hyperparameter retuning.

Table 9 summarizes the cross-site results. PVGS achieved 72.8% AP@0.5 and 60.45% AP@0.5:0.95 on Site B, corresponding to decreases of 1.50 and 2.47 percentage points relative to Site A. Precision and recall decreased by 1.50 and 2.20 points, respectively, while plant-level counting accuracy remained 98.4%, only 0.80 percentage points below the primary-site result. The limited degradation across the terrain, temperature, illumination, and layout shift provides direct evidence of cross-site transferability under the evaluated conditions.

Table 9. Cross-site generalization performance on the independent held-out desert photovoltaic plant (Site B).

Evaluation Metric	Primary Mountainous Site (Site A)	Held-Out Desert Site (Site B)	Delta/Drop
Detector AP@0.5 (%)	74.30	72.80	−1.50 pp
Detector AP@0.5:0.95 (%)	62.92	60.45	−2.47 pp
Precision (%)	92.00	90.50	−1.50 pp
Recall (%)	91.40	89.20	−2.20 pp
Plant-level counting accuracy (%)	99.20	98.40	−0.80 pp

Note: Site B was held out from Site A training and was evaluated using the Site-A-trained model without retraining or hyperparameter retuning. Performance differences are Site B minus Site A and are reported in percentage points (pp).

4.6. End-to-End Two-Pass Runtime

The runtime results cover density-estimation inference, final detector inference, image input/output and preprocessing, geographic projection, and DBSCAN aggregation (Table 10 and Figure 13). The five reported stage latencies total 15.1 ms/image, corresponding to 66.2 FPS by reciprocal conversion. Final detector inference contributes 9.5 ms/image, followed by the preliminary density-estimation pass at 3.2 ms/image. This

stage-wise accounting quantifies the processing cost of adaptive resolution and geographic aggregation. The rounded final-pass latency gives 105.3 FPS within this timing analysis. Table 1 reports 105.6 FPS for the primary final-pass benchmark, and Table 2 reports 9.9 ms/image for the separate detector-comparison experiment.

Table 10. Two-pass processing latency and reciprocal throughput.

Pipeline Component	Latency (ms/Image)	Equivalent Throughput (FPS)
Pass 1: density-estimation inference	3.2	312.5
Pass 2: final detector inference	9.5	105.3
Image I/O and preprocessing	1.1	909.1
Geospatial projection	0.9	1111.1
DBSCAN duplicate removal	0.4	2500.0
Total end-to-end pipeline	15.1	66.2

Note: Equivalent throughput equals 1000 divided by the listed latency in ms/image. The total aggregates the five processing stages. Results retain the scope of this runtime analysis; the detector-comparison results are reported separately in Table 2.

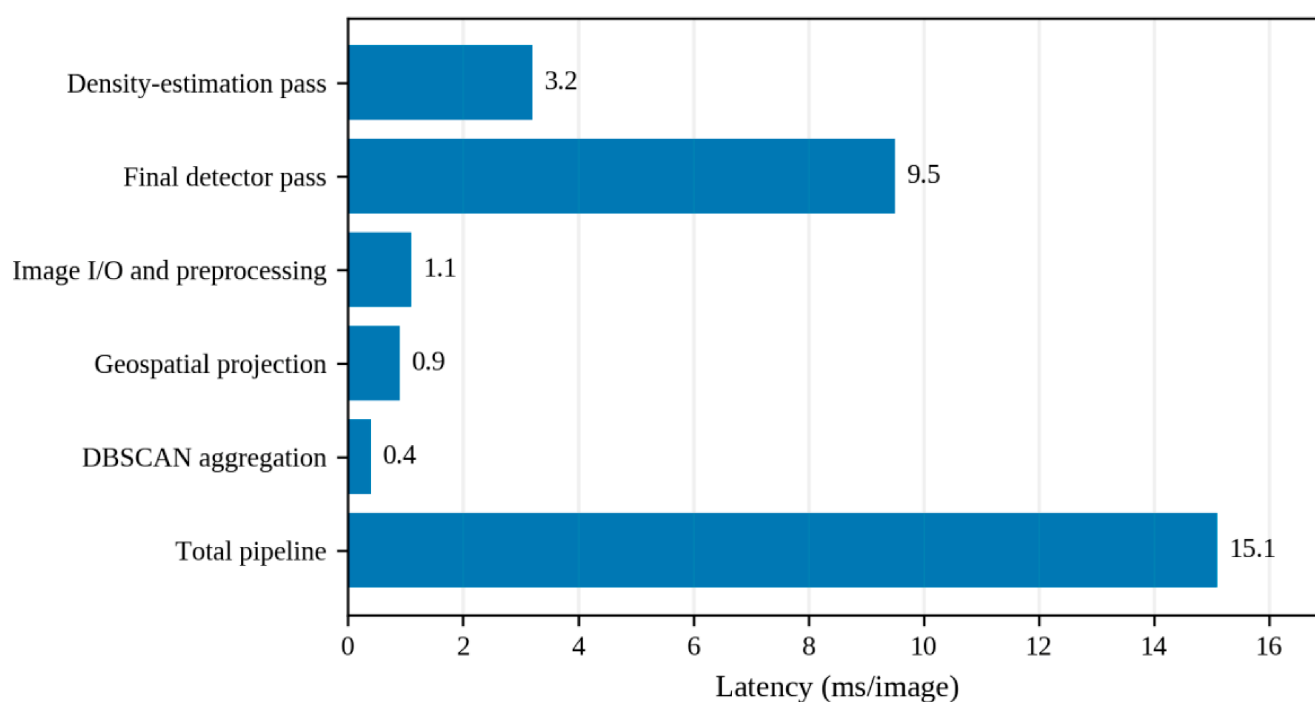


Figure 13. Latency of the inference and geographic aggregation stages in Table 10. The total pipeline bar represents the sum of the five preceding components.

4.7. Training-Set Size and Learning-Curve Analysis

To empirically examine how detector performance scales with training-set size, we trained the complete PVGS configuration using 20%, 40%, 60%, 80%, and 100% of the geographically blocked training partition while keeping the validation and test partitions fixed. AP@0.5 increased from 68.4% at 20% of the training data to 71.1%, 73.0%, 74.1%, and 74.3% at 40%, 60%, 80%, and 100%, respectively. The curve shows diminishing marginal gains beyond 80% of the available training data. Table 11 summarizes the training-data fractions, nominal image counts, and corresponding AP@0.5 values.

Table 11. Empirical learning-curve results for different fractions of the training partition.

Training Fraction (%)	Approx. Training Images	AP@0.5 (%)
20	672	68.4
40	1344	71.1
60	2016	73.0
80	2688	74.1
100	3360	74.3

Note: Image counts are nominal values based on the 80% training allocation of the 4200-image dataset (8:1:1 train/validation/test split). Geographic blocks remain indivisible, so exact block-constrained counts can differ slightly from the nominal arithmetic values. Validation and test partitions were held fixed and were not used for training.

The empirical curve indicates that approximately 80% of the available training partition (about 2688 nominal images) is sufficient to reach $AP@0.5 \approx 74\%$, only 0.2 percentage points below the full-data result. This estimate is specific to the present PVGS task and data distribution; it is not interpreted as a distribution-free minimum sample size. The result complements the theoretical sample-size perspective discussed in Ref. [32].

5. Discussion

5.1. Complementary Components and Engineering Implications

The primary repeated-run ablation and the component-level experiments support complementary roles for scene-targeted augmentation and detector modifications. In the primary experiment, the complete model improved $AP@0.5$ by 4.60 percentage points, $AP@0.5:0.95$ by 8.42 points, and recall by 6.20 points relative to the baseline. The isolated tests in Section 4.3 further show how RAM benefits reflection-affected imagery and how DFPN improves multi-scale detection. Because the experiments use distinct evaluation scopes, their numerical improvements are interpreted within each comparison, not summed into a single combined effect. Five-run statistics quantify training variability on the primary test set rather than the full range of deployment uncertainty.

The plant-level result is equally important for engineering use. The pose-constrained aggregation method achieved 99.2% counting accuracy, compared with 75.1% for the evaluated image-stitching workflow. In repetitive photovoltaic-array scenes, retaining the geographic identity of individual detections can be more reliable than constructing a visually continuous mosaic before counting. Because the output preserves geographic coordinates for the detected modules, it can support asset indexing, spatial quality control, targeted revisit planning, maintenance records, and comparison between repeated UAV inspections.

The radius $\varepsilon = D/2 + 2\sigma_{\text{err}}$ provides a geometric starting point for DBSCAN. With $D = 2.3$ m and $\sigma_{\text{err}} = 0.20$ m, $\varepsilon = 1.55$ m. The sensitivity experiment in Section 4.5.2 shows a stable net count over several neighboring radii at $\text{MinPts} = 2$, while also demonstrating that splitting and merging can offset each other. The regional experiment reinforces this distinction: 99.80% net accuracy coexists with 99.42% reference-count-weighted regional accuracy. Counts, geographic association, and regional error should therefore be assessed together when using the output for asset management.

The runtime analysis provides complementary detector-level and processing-chain results. Table 1 reports final-pass throughput, while Section 4.6 includes density estimation, image handling, projection, and clustering. Its five stage latencies total 15.1 ms/image, equivalent to 66.2 FPS. The results quantify the specified processing chain; acquisition, storage, and downstream inspection have separate timing boundaries.

5.2. Comprehensive Detector Benchmark on the Original Dataset

To provide a more comprehensive evaluation and address the need for broader model comparison, we implemented and evaluated an extended set of representative detectors, including SSD, YOLOv10n, YOLO11n, and MS-YOLOv11 [34], on the original curated UAV dataset. All models used the same real-world training and test partitions described in Section 3.4, enabling direct comparison under identical geometric and radiometric conditions. The benchmark results are summarized in Table 12, and Figure 14 displays the corresponding precision–recall operating points.

Table 12. Comprehensive detector benchmark conducted on the original UAV dataset.

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)
Faster R-CNN	85.24	88.40	89.15	62.30	41.53
SSD	82.10	79.55	81.30	55.45	24.30
YOLOv5n	88.35	86.12	89.50	70.15	1.90
YOLOv8n	90.12	89.45	92.30	75.80	3.20
YOLOv10n	92.50	91.20	95.80	82.10	2.30
YOLO11n	93.76	92.55	97.10	85.00	2.59
MS-YOLOv11	96.32	95.74	98.50	≈88.8	2.04

Note: All models were trained and evaluated from scratch on our unified primary UAV dataset. The evaluation metrics and hyperparameter configurations were strictly aligned with the primary PVGS experiments to ensure objective comparison, eliminating any data source discrepancies.

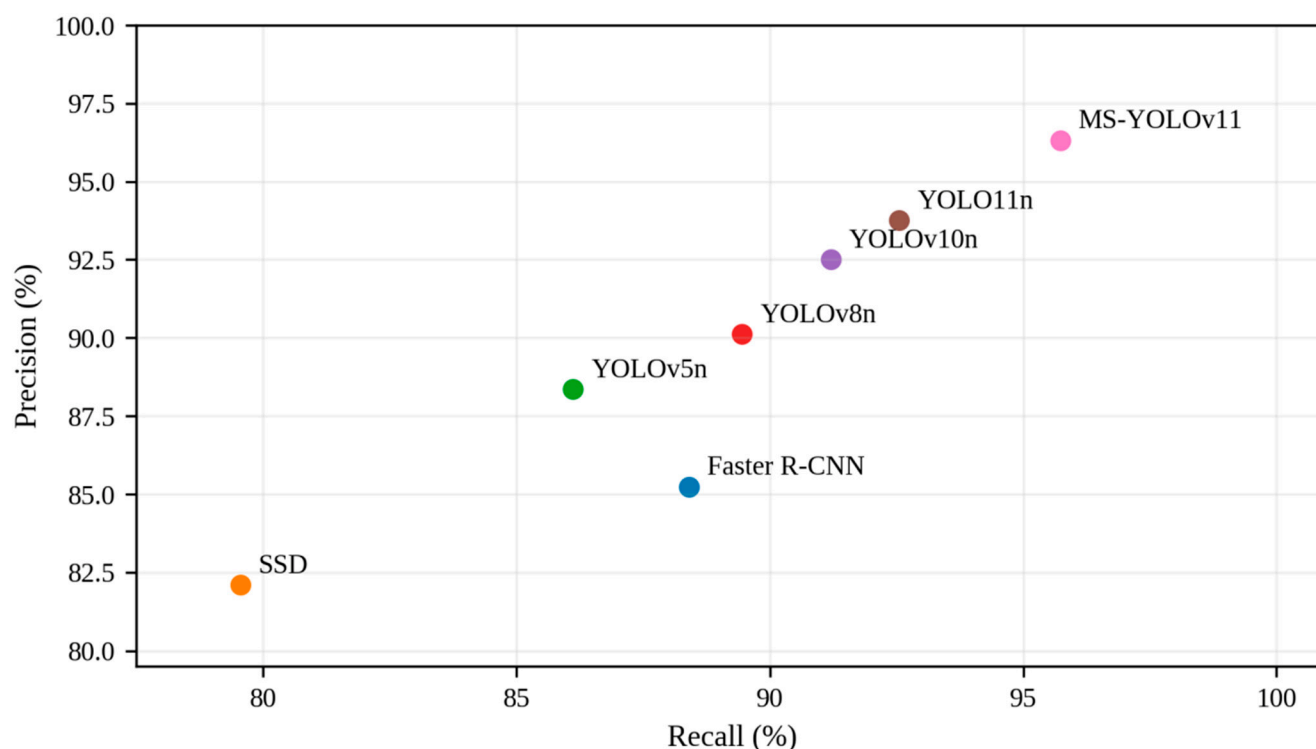


Figure 14. Precision–recall operating points for all seven detectors in Table 12, using the benchmark values. Each point is one reported operating point, not a complete PR curve.

5.3. Experimental Scope and Cross-Site Transferability

The primary plant-level experiment achieved 99.2% counting accuracy with the evaluated UAV imagery, pose records, camera geometry, image overlap, and local-plane surface representation. The independent Site B test extends this evaluation to a flat, high-temperature desert environment with intense glare and approximately 45,000 panel in-

stances. Without retuning the Site-A-trained model, PVGS retained 72.8% AP@0.5, 60.45% AP@0.5:0.95, 90.50% precision, 89.20% recall, and 98.4% plant-level counting accuracy. The relatively limited performance drop provides direct evidence that the detector and geospatial aggregation pipeline transfer beyond the primary mountainous site under the evaluated cross-site shift.

PV-PhysAug was evaluated through the operator outputs in Figure 4, the primary ablation in Section 4.2, the operator-wise ablation in Section 4.3.3, and the quantitative plausibility analysis in Section 4.3.5. The complete augmentation strategy improved AP@0.5 by 1.92 percentage points relative to the baseline, while the component study confirmed positive contributions from reflection, occlusion, perspective, and boundary perturbations. The validation-image comparison yielded SSIM = 0.842 ± 0.035 and FID = 14.68. Combined with the held-out Site B evaluation, these results provide complementary evidence that the physics-guided augmentation and adaptive resolution mechanisms are not limited to a single acquisition environment. Direct radiometric calibration remains a separate validation task. The empirical learning-curve experiment in Section 4.7 showed diminishing returns beyond 80% of the training partition and provides a task-specific sample-size estimate, complementing the theoretical learning-curve perspective in Ref. [32].

The four-layout evaluation covers flat utility-scale rows, moderately tilted rows, strong glare, and heterogeneous panel spacing within the original dataset, while the independent Site B experiment adds a geographic and environmental domain shift. Together, these experiments substantially strengthen the generalization assessment. Further multi-site validation across additional UAV platforms, flight heights, seasons, and panel technologies would broaden the deployment envelope. Explicit geolocation uncertainty, kinematic positioning, higher-fidelity surface models, and uncertainty-aware clustering are additional development directions. Combining visible and thermal imagery could link the georeferenced inventory with condition and defect information.

The experimental summaries provide complementary levels of statistical detail. The five-run ablation includes individual observations, confidence intervals, and paired comparisons. The component and attention tables provide their reported means and standard deviations, while the remaining comparisons retain their tabulated metrics and timing scopes. ROC interpretation uses the specified positive and negative sample definitions. Each set of results is interpreted according to its evaluation population and reporting detail.

Methodological reproducibility and data accessibility are treated separately. The UAV imagery and site metadata remain subject to third-party restrictions; however, Algorithms 1–3 provide pseudocode for the augmentation, detector-inference, and geospatial-aggregation procedures, Section 3.2 specifies the RAM/DFPN layer operations, and Section 3.4 provides the complete training hyperparameters and random seeds. Supplementary Code S1 provides the core implementation used to reproduce the method without distributing the proprietary imagery or site-identifying metadata.

6. Conclusions

This study presents an integrated UAV-based workflow that links photovoltaic panel detection with duplicate-free plant-level aggregation. The detector combines scene-targeted augmentation, adaptive resolution, channel attention, and decoupled multi-scale fusion, while the post-detection stage uses camera pose and spatial clustering to transform overlapping observations into unique geographic assets.

Across five repeated runs, the complete detector achieved 74.3% AP@0.5 and 62.9% AP@0.5:0.95. Pose-assisted aggregation achieved 99.2% normalized counting accuracy in the primary plant-level experiment, compared with 75.1% for image stitching. On the independent held-out desert Site B, the Site-A-trained model achieved 72.8% AP@0.5, 60.45%

AP@0.5:0.95, 90.50% precision, 89.20% recall, and 98.4% plant-level counting accuracy without retuning, providing direct cross-site generalization evidence. Isolated RAM and DFPN tests, attention comparisons, and PR/ROC analyses provide complementary detector-level evidence within their respective evaluation scopes.

The regional counting experiment yielded 99.80% net accuracy and 99.42% reference-count-weighted regional accuracy, demonstrating the importance of distinguishing total error from regional error compensation. Across four evaluated photovoltaic layouts, counting accuracy ranged from 98.91% to 99.38%. The two-pass runtime experiment reported 15.1 ms/image for the specified processing chain, equivalent to 66.2 FPS, while 105.6 FPS describes the primary final detector pass. Together with the held-out Site B evaluation, these results support the integrated detection-and-inventory workflow across the evaluated imaging conditions, layouts, and geographic settings. The operator-wise augmentation experiment and density-stratified adaptive-resolution experiment further confirmed that the gains arise from complementary perturbations and that resolution scaling is most beneficial in dense/small-target scenes.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app16199719/s1>: Code S1: Core Python Implementation of PVGS for Photovoltaic Panel Detection and Geospatial Aggregation.

Author Contributions: Conceptualization, S.Z.; methodology, X.Z.; validation, Y.H.; data curation, Y.H.; writing—original draft preparation, K.L.; writing—review and editing, X.Z.; supervision, H.H.; funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Research Project of Henan Province, grant number 262102321110, and the National Natural Science Foundation of China, grant number 42504128.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the UAV imagery, georeferenced metadata, and complete primary-site inventory used in this study. These materials were obtained under third-party access conditions and cannot be publicly distributed without permission. Requests for data access should be directed to the corresponding author and are subject to the data provider's approval. The manuscript reports the experimental summaries and the designated regional inventory in Table 7. Algorithms 1–3 provide implementation-level pseudocode, Section 3.2 specifies the RAM/DFPN architecture, and Section 3.4 provides the complete training hyperparameters. Core implementation files are supplied as Supplementary Code S1; the archive does not contain proprietary UAV imagery, coordinates, or site-identifying metadata.

Acknowledgments: The authors thank Wenxin Feng for assistance with software implementation.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Ablation and Statistical Details

Appendix A preserves the full repeated-run evidence underlying the summary in Section 4.2 and the detailed component-level comparisons summarized in Section 4.3. Table A1 reports all individual run values, Figure A1 visualizes the AP@0.5 distribution across the five runs, Table A2 gives aggregated mean $\pm$ SD values and 95% confidence intervals, Table A3 reports paired statistical comparisons against the YOLO11 baseline, and Tables A4 and A5 provide the isolated RAM/DFPN and attention-module results.

Table A1. Individual results for the five repeated-run ablation experiments.

Run	Configuration	AP@0.5 (%)	AP@0.5:0.95 (%)	Recall (%)	Final-Pass FPS
Run 1	Baseline	69.5	54.1	84.8	112.3
Run 1	Baseline + PV-PhysAug	71.2	56.5	87.0	110.8
Run 1	Baseline + RAM + DFPN	71.9	59.3	88.3	104.7
Run 1	Baseline + Adaptive resolution	70.2	55.5	86.3	98.4
Run 1	PVGS (complete)	73.9	62.5	91.0	105.6
Run 2	Baseline	69.9	54.9	85.5	112.3
Run 2	Baseline + PV-PhysAug	71.9	57.3	87.9	110.8
Run 2	Baseline + RAM + DFPN	72.6	60.1	89.2	104.7
Run 2	Baseline + Adaptive resolution	70.6	56.0	86.9	98.4
Run 2	PVGS (complete)	74.4	63.0	91.7	105.6
Run 3	Baseline	69.8	54.7	85.0	112.3
Run 3	Baseline + PV-PhysAug	71.5	57.0	87.4	110.8
Run 3	Baseline + RAM + DFPN	72.5	60.0	88.8	104.7
Run 3	Baseline + Adaptive resolution	70.3	55.7	86.5	98.4
Run 3	PVGS (complete)	74.5	63.2	91.5	105.6
Run 4	Baseline	69.6	54.3	85.3	112.3
Run 4	Baseline + PV-PhysAug	71.7	56.8	87.6	110.8
Run 4	Baseline + RAM + DFPN	72.0	59.5	88.5	104.7
Run 4	Baseline + Adaptive resolution	70.5	55.9	87.0	98.4
Run 4	PVGS (complete)	74.1	62.8	91.2	105.6
Run 5	Baseline	69.7	54.5	85.4	112.3
Run 5	Baseline + PV-PhysAug	71.8	57.4	87.7	110.8
Run 5	Baseline + RAM + DFPN	72.8	60.4	89.5	104.7
Run 5	Baseline + Adaptive resolution	70.4	56.2	86.7	98.4
Run 5	PVGS (complete)	74.6	63.1	91.6	105.6
Mean $\pm$ SD	Baseline	69.70 $\pm$ 0.16	54.50 $\pm$ 0.32	85.20 $\pm$ 0.29	112.3
Mean $\pm$ SD	Baseline + PV-PhysAug	71.62 $\pm$ 0.28	57.00 $\pm$ 0.37	87.52 $\pm$ 0.34	110.8
Mean $\pm$ SD	Baseline + RAM + DFPN	72.36 $\pm$ 0.39	59.86 $\pm$ 0.45	88.86 $\pm$ 0.49	104.7
Mean $\pm$ SD	Baseline + Adaptive resolution	70.40 $\pm$ 0.16	55.86 $\pm$ 0.27	86.68 $\pm$ 0.29	98.4
Mean $\pm$ SD	PVGS (complete)	74.30 $\pm$ 0.29	62.92 $\pm$ 0.28	91.40 $\pm$ 0.29	105.6

Note: Runs 1–5 correspond to seeds 3407, 2024, 2025, 2026, and 4561334, respectively. Final-pass throughput is listed alongside the corresponding accuracy run.

Table A2. Aggregated five-run ablation results (mean $\pm$ SD with 95% confidence intervals).

Configuration	AP@0.5 (%)	AP@0.5:0.95 (%)	Recall (%)	Final-Pass FPS
YOLO11 baseline	69.70 $\pm$ 0.16 [69.50, 69.90]	54.50 $\pm$ 0.32 [54.11, 54.89]	85.20 $\pm$ 0.29 [84.84, 85.56]	112.3
Baseline + PV-PhysAug	71.62 $\pm$ 0.28 [71.28, 71.96]	57.00 $\pm$ 0.37 [56.54, 57.46]	87.52 $\pm$ 0.34 [87.10, 87.94]	110.8
Baseline + RAM + DFPN	72.36 $\pm$ 0.39 [71.87, 72.85]	59.86 $\pm$ 0.45 [59.30, 60.42]	88.86 $\pm$ 0.49 [88.25, 89.47]	104.7
Baseline + Adaptive resolution	70.40 $\pm$ 0.16 [70.20, 70.60]	55.86 $\pm$ 0.27 [55.52, 56.20]	86.68 $\pm$ 0.29 [86.32, 87.04]	98.4
PVGS (complete)	74.30 $\pm$ 0.29 [73.94, 74.66]	62.92 $\pm$ 0.28 [62.58, 63.26]	91.40 $\pm$ 0.29 [91.04, 91.76]	105.6

Note: Values in square brackets are 95% t-based confidence intervals. Final-pass FPS is a configuration-level throughput value and is shown without a standard deviation.



Figure A1. Distribution of AP@0.5 across five repeated training runs for the ablation configurations. Boxes show the median and interquartile range; points show the five individual runs.

Table A3. Paired statistical comparisons against the YOLO11 baseline.

Configuration	Metric	Mean Difference (pp)	95% CI (pp)	<i>p</i> -Value	Cohen's <i>d_z</i>
Baseline + PV-PhysAug	AP@0.5	1.92	[1.67, 2.17]	3.069×10^{-5}	9.37
Baseline + PV-PhysAug	AP@0.5:0.95	2.50	[2.21, 2.79]	1.837×10^{-5}	10.66
Baseline + PV-PhysAug	Recall	2.32	[2.22, 2.42]	4.052×10^{-7}	27.73
Baseline + RAM + DFPN	AP@0.5	2.66	[2.30, 3.02]	3.251×10^{-5}	9.23
Baseline + RAM + DFPN	AP@0.5:0.95	5.36	[4.98, 5.74]	2.504×10^{-6}	17.58
Baseline + RAM + DFPN	Recall	3.66	[3.24, 4.08]	1.689×10^{-5}	10.89
Baseline + Adaptive resolution	AP@0.5	0.70	[0.52, 0.88]	3.790×10^{-4}	4.95
Baseline + Adaptive resolution	AP@0.5:0.95	1.36	[0.98, 1.74]	5.681×10^{-4}	4.46
Baseline + Adaptive resolution	Recall	1.48	[1.30, 1.66]	2.389×10^{-5}	9.98
PVGS (complete)	AP@0.5	4.60	[4.35, 4.85]	8.555×10^{-7}	23.00
PVGS (complete)	AP@0.5:0.95	8.42	[8.18, 8.66]	6.532×10^{-8}	43.77
PVGS (complete)	Recall	6.20	[5.94, 6.46]	3.284×10^{-7}	29.23

Note: Mean differences and 95% confidence intervals are in percentage points. *p*-values are two-sided paired *t*-test results; Cohen's *d_z* is the paired standardized effect size. The exact two-sided Wilcoxon signed-rank *p*-value is 0.0625 for each all-positive five-pair comparison.

The component-level RAM/DFPN and attention-module comparison tables are also retained in this appendix to keep Section 4 concise while preserving the complete numerical evidence requested during review.

Table A4. Isolated RAM and DFPN ablation across photovoltaic image subsets.

Configuration	Subset	AP@0.5 (%)	AP@0.5:0.95 (%)	Precision (%)	Recall (%)
YOLO11 baseline	All	94.63 ± 0.39	82.73 ± 0.39	81.03 ± 0.89	87.92 ± 0.34
YOLO11 baseline	Reflection-affected	84.75 ± 1.40	72.85 ± 1.40	72.51 ± 1.07	76.42 ± 1.69
YOLO11 baseline	Normal	98.19 ± 0.34	86.29 ± 0.34	85.03 ± 0.88	92.46 ± 0.92
YOLO11 + RAM	All	97.31 ± 0.16	85.41 ± 0.16	85.29 ± 0.46	90.71 ± 0.55
YOLO11 + RAM	Reflection-affected	93.92 ± 0.72	82.02 ± 0.72	81.42 ± 1.24	86.55 ± 1.09
YOLO11 + RAM	Normal	98.45 ± 0.30	86.55 ± 0.30	85.78 ± 0.46	93.83 ± 0.59
YOLO11 + DFPN	All	96.68 ± 0.29	84.78 ± 0.29	84.47 ± 0.74	90.11 ± 0.54
YOLO11 + DFPN	Reflection-affected	88.72 ± 0.45	76.82 ± 0.45	75.30 ± 0.94	81.00 ± 0.70
YOLO11 + DFPN	Normal	99.00 ± 0.40	87.42 ± 0.40	87.23 ± 0.99	94.46 ± 0.74
YOLO11 + RAM + DFPN	All	99.00 ± 0.18	87.45 ± 0.18	89.50 ± 0.80	93.54 ± 0.64
YOLO11 + RAM + DFPN	Reflection-affected	97.30 ± 0.26	85.40 ± 0.26	85.57 ± 0.63	90.52 ± 0.72
YOLO11 + RAM + DFPN	Normal	99.00 ± 0.19	88.25 ± 0.19	90.44 ± 0.75	95.74 ± 0.58

Note: Entries give the reported mean ± SD for the component-level experiment. “All” denotes the complete evaluation set used for this component comparison. Other experiment series retain their own evaluation sets.

Table A5. Attention-module comparison on the reflection-affected subset.

Attention Module	AP@0.5 (%)	AP@0.5:0.95 (%)	Precision (%)	Recall (%)
SE	91.02 ± 0.38	79.12 ± 0.38	78.58 ± 0.82	82.95 ± 0.77
CBAM	92.57 ± 0.40	80.67 ± 0.40	80.61 ± 0.66	84.77 ± 1.49
RAM (proposed)	94.42 ± 0.41	82.52 ± 0.41	81.53 ± 0.64	87.53 ± 1.14

Note: Values are the reported mean ± SD for the attention-comparison experiment. The RAM result here and the RAM result in Table A4 refer to their respective experimental series.

References

- International Energy Agency. *World Energy Outlook*; OECD/IEA: Paris, France, 2009.
- Starzyński, J.; Zawadzki, P.; Harańczyk, D. Machine learning in solar plants inspection automation. *Energies* **2022**, *15*, 5966. [\[CrossRef\]](#)
- Polymeropoulos, I.; Bezyrgiannidis, S.; Vrochidou, E.; Papakostas, G.A. Enhancing solar plant efficiency: A review of vision-based monitoring and fault detection techniques. *Technologies* **2024**, *12*, 175. [\[CrossRef\]](#)
- Abro, G.E.M.; Ali, A.; Memon, S.A.; Memon, T.D.; Khan, F. Strategies and challenges for unmanned aerial vehicle-based continuous inspection and predictive maintenance of solar modules. *IEEE Access* **2024**, *12*, 176615–176629. [\[CrossRef\]](#)
- Vega Díaz, J.J.; Vlamincx, M.; Lefkaditis, D.; Vargas, S.A.O.; Luong, H. Solar panel detection within complex backgrounds using thermal images acquired by UAVs. *Sensors* **2020**, *20*, 6219. [\[CrossRef\]](#) [\[PubMed\]](#)
- Barraz, Z.; Sebari, I.; Lamrini, N.; El Kadi, K.A.; Abdelmoula, I.A. Fast and automatic solar module geo-labeling for optimized large-scale photovoltaic systems inspection from UAV thermal imagery using deep learning segmentation. *Clean. Eng. Technol.* **2025**, *28*, 101048. [\[CrossRef\]](#)
- Su, Z.; Zhao, K.; Fan, Z.; Guo, X. RTPV-YOLO: Real-time photovoltaic detection with UAV-based thermal and RGB imaging. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 13799–13810. [\[CrossRef\]](#)
- Rodriguez-Vazquez, J.; Prieto-Centeno, I.; Fernandez-Cortizas, M.; Perez-Saura, D.; Molina, M.; Campoy, P. Real-time object detection for autonomous solar farm inspection via UAVs. *Sensors* **2024**, *24*, 777. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jumaboev, S.; Jurakuziev, D.; Lee, M. Photovoltaics plant fault detection using deep learning techniques. *Remote Sens.* **2022**, *14*, 3728. [\[CrossRef\]](#)
- Saeed, F.; Ahmed, M.J.; Gul, M.J.; Hong, K.J.; Paul, A.; Kavitha, M.S. A robust approach for industrial small-object detection using an improved faster regional convolutional neural network. *Sci. Rep.* **2021**, *11*, 23390. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hussain, T.; Hussain, M.; Al-Aqrabi, H.; Alsabui, T.; Hill, R. A review on defect detection of electroluminescence-based photovoltaic cell surface images using computer vision. *Energies* **2023**, *16*, 4012. [\[CrossRef\]](#)
- Li, J.; Tong, T.; Li, D.; Yuan, X.; Liu, P.; Zhang, J.; Zhu, X.; Zhao, D.; Fang, H. A dynamically adaptive and high-efficiency small object detection network for infrared thermographic images in online monitoring of solar photovoltaic panel defects. *Energy* **2025**, *335*, 138129. [\[CrossRef\]](#)

13. Arnaudo, E.; Blanco, G.; Monti, A.; Bianco, G.; Monaco, C.; Pasquali, P.; Dominici, F. A comparative evaluation of deep learning techniques for photovoltaic panel detection from aerial images. *IEEE Access* **2023**, *11*, 47579–47594. [\[CrossRef\]](#)
14. Rahman, M.R.U.; Chen, H. Defects inspection in polycrystalline solar cells electroluminescence images using deep learning. *IEEE Access* **2020**, *8*, 40547–40558. [\[CrossRef\]](#)
15. Zhang, J.; Yang, W.; Chen, Y.; Ding, M.; Huang, H.; Wang, B.; Gao, K.; Chen, S.; Du, R. Fast object detection of anomaly photovoltaic cells using deep neural networks. *Appl. Energy* **2024**, *372*, 123759. [\[CrossRef\]](#)
16. Chen, H.; Song, M.; Zhang, Z.; Liu, K. Detection of surface defects in solar cells by bidirectional-path feature pyramid group-wise attention detector. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5025609. [\[CrossRef\]](#)
17. Hijjawi, U.; Lakshminarayana, S.; Xu, T.; Fierro, G.P.M.; Rahman, M. A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations. *Sol. Energy* **2023**, *266*, 112186. [\[CrossRef\]](#)
18. Ma, T.; Wang, H.; Liang, J.; Wang, Y.; Peng, J.; Kai, Z.; Liu, X. The temporal-spatial information fusion network for multi-frame infrared small target detection. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 4505219. [\[CrossRef\]](#)
19. Pierdicca, R.; Paolanti, M.; Felicetti, A.; Piccinini, F.; Zingaretti, P. Automatic faults detection of photovoltaic farms: SolAIr, a deep learning-based system for thermal images. *Energies* **2020**, *13*, 6496. [\[CrossRef\]](#)
20. Sun, Z.; Meng, C.; Huo, L.; Li, T.; Zhang, Z.; Chang, S. An embedded feature pyramid network enables bidirectional information flow for object detection and instance segmentation. *Neurocomputing* **2025**, *619*, 129196. [\[CrossRef\]](#)
21. Li, Y.; Liu, F.; Qin, Y.; Dai, L.; Zhang, L.; Xie, W. Small object detection network based on progressive enhanced multi-level feature fusion. *Neurocomputing* **2025**, *649*, 130755. [\[CrossRef\]](#)
22. Yuan, Z.; Gong, J.; Guo, B.; Wang, C.; Liao, N.; Song, J.; Wu, Q. Small object detection in UAV remote sensing images based on intra-group multi-scale fusion attention and adaptive weighted feature fusion mechanism. *Remote Sens.* **2024**, *16*, 4265. [\[CrossRef\]](#)
23. Costa, M.V.C.V.; Carvalho, O.L.F.; Orlandi, A.G.; Hirata, I.; Albuquerque, A.O.d.; Silva, F.V.e.; Guimarães, R.F.; Gomes, R.A.T.; Júnior, O.A.d.C. Remote sensing for monitoring photovoltaic solar plants in Brazil using deep semantic segmentation. *Energies* **2021**, *14*, 2960. [\[CrossRef\]](#)
24. Ravishankar, R.; AlMahmoud, E.; Habib, A.; de Weck, O.L. Capacity estimation of solar farms using deep learning on high-resolution satellite imagery. *Remote Sens.* **2023**, *15*, 210. [\[CrossRef\]](#)
25. Liu, W.; Huo, H.; Ji, L.; Zhao, Y.; Liu, X.; Li, J. A method for extracting photovoltaic panels from high-resolution optical remote sensing images guided by prior knowledge. *Remote Sens.* **2024**, *16*, 9. [\[CrossRef\]](#)
26. Zhang, H.; Tian, P.; Zhong, J.; Liu, Y.; Li, J. Mapping photovoltaic panels in coastal China using Sentinel-1 and Sentinel-2 images and Google Earth Engine. *Remote Sens.* **2023**, *15*, 3712. [\[CrossRef\]](#)
27. Zhu, A.; Xie, J.; Wang, B.; Guo, H.; Guo, Z.; Wang, J.; Xu, L.; Zhu, S.; Yang, Z. Lightweight defect detection algorithm of tunnel lining based on knowledge distillation. *Sci. Rep.* **2024**, *14*, 27178. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Zhu, A.; Xie, J.; Guo, H.; Wang, J.; Guo, Z.; Xu, L.; Zhu, S.; Yang, Z.; Wang, B. Surface defect detection model of laser cutting polycrystalline cubic boron nitride tool based on asymptotic fusion strategy. *Sci. Rep.* **2024**, *14*, 26689. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Meng, F.; Jiang, T.; Zhu, S.; Wan, L. Dual seismic image collaborative recognition algorithm based on deep learning. *Eng. Appl. Artif. Intell.* **2025**, *162*, 112800. [\[CrossRef\]](#)
30. Zhou, Y.; Wei, Y. UAV-DETR: An Enhanced RT-DETR Architecture for Efficient Small Object Detection in UAV Imagery. *Sensors* **2025**, *25*, 4582. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Guo, H.; Wu, Q.; Wang, Y. AUHF-DETR: A Lightweight Transformer with Spatial Attention and Wavelet Convolution for Embedded UAV Small Object Detection. *Remote Sens.* **2025**, *17*, 1920. [\[CrossRef\]](#)
32. Salazar, A.; Vergara, L.; Vidal, E. A proxy learning curve for the Bayes classifier. *Pattern Recognit.* **2023**, *136*, 109240. [\[CrossRef\]](#)
33. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [\[CrossRef\]](#)
34. Zhu, J.; Jin, H.; Gao, D.; Mei, X.; Chen, S.; Qiu, J.; Zhang, H. MS-YOLOv11: A multi-scale feature fusion network for real-time rooftop photovoltaic detection from UAV images. *PLoS ONE* **2026**, *21*, e0346424. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.