# Phantom Fires: Safe False-Alert Suppression via Multimodal Fusion of a Visible-Patch CNN and Solar Reflection Geometry for Thermal Fault Monitoring

Shuaib Hanief[1], Ghulam Jilani Raza[1], David Halnon[1], Rana Shujaat Ali[1], and Abhishek Madaan[1]

[1] Multisensor AI, Inc., Houston, TX, United States
shuaib.hanief@multisensorai.com
ghulam.raza@multisensorai.com
david.halnon@multisensorai.com
shujaat.ali@multisensorai.com
abhishek.madaan@multisensorai.com

Abstract

Continuous outdoor thermal monitoring raises an analyst alert whenever an IR pixel exceeds a fixed apparent-temperature threshold. We report a 40-day deployment over 10 cameras (4,143.6 camera-hours) that produced 157 above-threshold events ($\geq$ 150 °C): 122 (77.7 %) solar-glare false positives and 35 (22.3 %) confirmed true thermal events — a glare event rate of 0.0294 events per camera-hour (95 % CI [0.0157, 0.0449], bootstrap $n = 1000$). To suppress this false-alarm burden without compromising thermal-event recall, we construct a paired visible-patch dataset of 10,622 RGB patches and evaluate a ladder of five glare/noglare discriminators of increasing complexity, from a brightness threshold through a forward Sandia SGHAT geometric model, a logistic regression over 28 hand-crafted features, a small CNN (245,889 parameters), and a multimodal late fusion of the CNN and geometric scores. The fusion classifier reaches 98.6 % accuracy on a balanced 2,126-patch held-out eval set, a 3.3-percentage-point gain over the CNN alone and a 14.1-point gain over the geometric model; the learned weights ($w_{\text{cnn}} = 7.34$, $w_{\text{geo}} = 4.50$, bias $= -1.93$) confirm both streams carry independent signal, and adding the geometric stream cuts the CNN's false-positive count by 73 % while preserving recall. A separate system-level test on the 35 in-window true thermal events shows that both the CNN and the fusion classifier reach 99.58 % specificity (235 of 236 IR/VIS frame pairs preserved at the 150 °C production threshold; 34 / 35 events preserved, event-level 95 % CI [85.1 %, 99.9 %]). The current production deployment runs no glare suppression — every above-threshold IR event raises an analyst alert today — so the five-rung ladder and the multimodal fusion are research-grade results, not deployed artefacts. We propose a three-way routing policy (alert, fallback, suppress) that wraps the best rung and report the measured anchors that constrain its expected outcome on the deployment ledger.



## 1. Introduction

Continuous outdoor thermal monitoring is increasingly used as a real-time safety system: fixed cameras stream infrared (IR) and visible imagery, and any pixel exceeding an apparent-temperature threshold raises an alert that an analyst must triage. Each such alert falls into one of two operational classes: a true thermal event that must reach the analyst, or a solar-glare false alert that should be suppressed before reaching the queue. The job of the routing system is therefore false-alert suppression with a strict safety constraint: glare suppressed, true thermal events preserved. Figure 1 illustrates the core failure mode: a specular reflection on a solar-panel array reads to the IR sensor as a bright high-temperature region (left), indistinguishable in the IR-only signal from a real thermal anomaly — the paired visible-light view (right) reveals the same scene as ordinary panel glare.

The deployment described here ran over a 40-day calendar window (2026-04-20 to 2026-05-30) across 10 cameras, accumulating 4,143.6 camera-hours of coverage (172.7 camera-days when summed across all cameras). Over that window, the system raised 157 above-threshold events ($\geq$ 150 °C). Domain review classified 122 of them (77.7 %) as solar-glare false positives reflecting off nearby solar-panel arrays and 35 (22.3 %) as


Figure 1. The "phantom fire" pattern motivating this paper. Left: long-wave IR image of a solar-panel array showing a bright, high-temperature region (false alert). Right: the paired visible image of the same scene at the same instant, in which the bright region is plainly the sun specularly reflected off the panel surface — not a thermal anomaly. Without a visible-modality check, the IR pixel-threshold rule cannot distinguish these two cases.

confirmed true thermal events. The operationally interesting quantity is the glare event rate per camera-hour — the burden the routing system is asked to suppress:

> 122 of 157 above-threshold events in this deployment were solar-glare false positives; the remaining 35 were confirmed true thermal events. Glare event rate: 0.0294 events per camera-hour, 95 % CI [0.0157, 0.0449].

Operationally, this volume of solar-glare false alerts imposes a significant cognitive burden on the on-site triage team and reduces overall trust in the deployed monitoring system: every alert in the queue must be inspected, and a high false-alert rate desensitises analysts to alerts in general — including the real thermal events the system exists to catch.

This paper reports an industry-experience study of how to suppress that false-alarm burden without compromising the system's ability to deliver a real thermal event when one occurs. The 35 confirmed true thermal events in the 40-day window are concentrated on Camera #2 (27) and Camera #3 (8), separate from the two cameras that produced the glare burden (Camera #1: 101 glare events; Camera #7: 21). The end-to-end test of the system on those 35 events is reported in §5.5: at the 150 °C production threshold, the 236 IR/VIS frame pairs that comprise these events are correctly preserved 235 / 236 times by both the CNN (Rung 4) and the multimodal fusion (Rung 5), a 99.58 % specificity. §6.5 reports the measured anchors that constrain the expected outcome of the §6 routing policy on the §3 ledger, and states explicitly why a per-event suppress / fallback / alert breakdown is not derivable from offline scores alone. Four contributions:

1. Field characterisation. We construct an event-level ledger of the deployment, expose the camera/time-of-day structure of the glare burden, and build a paired visible-patch dataset of 10,622 60 × 60 RGB patches (5,311 glare drawn from the two cameras that produced glare events, Camera #1 and Camera #7; 5,311 noglare drawn from a temporally disjoint pool spanning all 10 cameras) that supports binary classifier development.
2. A ladder of five discriminators, single-modality through multimodal. We evaluate five glare/noglare classifiers of increasing complexity: a brightness threshold (visible only), a forward SGHAT geometric model with Monte Carlo over pose uncertainty (geometry only), a logistic regression over 28 visible-patch features (visible only), a small CNN (visible only), and a multimodal late fusion of CNN and geometric scores. Each is reported with confusion matrix, PR curve, and operating-threshold sweep.
3. A proposed multimodal routing architecture. We define how the multimodal classifier output would integrate with the analyst alert queue via a conservative three-way decision policy (alert / fallback to human / suppress), with explicit roles for the geometric prior, the visible-modality cross-check, and human review of uncertain events. The current production deployment has no glare suppression in place — every above-threshold IR event becomes an analyst alert today — so this routing architecture is presented as the next deployment milestone, not as a description of what is in production.
4. Multimodal fusion outperforms either modality alone. On the held-out evaluation split, the visual-modality CNN (Rung 4) reaches 95.3 % accuracy and the geometric model alone (Rung 2, at its optimal threshold) reaches 84.5 %. The multimodal late fusion (Rung 5) reaches 98.6 % — a 3.3-percentage-point gain over the CNN and a 14.1-point gain over the geometric model. The fusion's learned weights ($w_{\text{cnn}} = 7.34$, $w_{\text{geo}} = 4.50$) confirm that both streams carry independent signal. This is the principal deployment justification for the multimodal approach: vision and physics are complementary modalities, and combining them buys real accuracy.

Section 2 positions the work; Section 3 quantifies the glare burden; Section 4 walks through each rung; Section 5 compares their classifier performance and reports the §5.5 system-level test-set evaluation on the 35 in-window true thermal events; Section 6 describes the proposed multimodal routing architecture that would wrap

the best rung at deployment (Section 6.4 explicitly contrasts this proposal against today's production setup, which runs no glare suppression) and reports the measured anchors that constrain its expected outcome on the §3 ledger (§6.5); Sections 7 and 8 discuss limits and outlook.

## 2. Background

Outdoor IR thermography of solar-relevant installations is governed by IEC TS 62446-3:2017, which mandates paired visible documentation alongside IR captures, plus controlled treatment of emissivity, reflected temperature, irradiance and wind (International Electrotechnical Commission, 2017). The IEA PVPS Task 13 report extends this with field-practice guidance for IR and electroluminescence (EL) imaging and flags reflections from buildings, panels and trees as confounders (Jahn et al., 2018). Both standards assume periodic, operator-controlled inspection. Our setting differs: the cameras are fixed, the acquisition is continuous, and the operational output is an alarm queue rather than an inspection report.

Glare physics is a mature area. Sandia's Solar Glare Hazard Analysis Tool (SGHAT) predicts whether sunlight reflected from a panel array reaches a known observer at a given timestamp (Ho, Sims, Yellowhair, & Wendelin, 2018; Ho et al., 2018). Hess and Bucher characterise photovoltaic (PV) module glare photographically and confirm its temporally repeatable, geometrically determined character (Hess & Bucher, 2024). We use SGHAT's vector-reflection equations directly, in their original forward direction. Our only extension is to wrap the deterministic test in a Monte Carlo over surveyed camera-position and panel-orientation uncertainty so that the output is a calibrated [0, 1] glare probability rather than a deterministic yes/no, suitable for use as a feature in the Rung 5 fusion classifier.

Visual classification of glare has been studied less. Hijjawi et al. review automated PV defect detection across imaging modalities (Hijjawi, Lakshminarayana, Xu, Malfense Fierro, & Rahman, 2023); Mellit and Kalogirou survey AI/IoT methods for PV diagnosis (Mellit & Kalogirou, 2021); Cardoso et al. fuse unmanned-aerial-vehicle (UAV) RGB+thermal data for module-level analysis (Cardoso, Jurado-Rodríguez, López, Ramos, & Jurado, 2024). None of these address glare specifically as a binary classification target on co-registered IR/visible patches in a fixed-camera fault-monitoring workflow.

The gap this paper addresses is therefore narrow and operational: deployment-scale event-level quantification of glare false alarms, plus an honest head-to-head of physics-only, feature-engineered, learned, and fused glare discriminators on a single labelled patch dataset.

## 3. The Glare Burden in a Real Deployment

### 3.1. Deployment and Hardware

The deployment site hosts 10 fixed-mount outdoor camera pairs (Summit-Series XTW model family) overlooking an industrial scene that includes a solar-panel array. Each camera pair captures a synchronised long-wave IR radiometric JPEG and an RGB visible JPEG; the two sensors are factory-aligned and per-camera $3 \times 3$ homographies are pre-calibrated for IR-to-visible registration. The analysis window spans 2026-04-20 to 2026-05-30, a 40-day calendar window. Per-camera coverage varies (five cameras cover the full window; five came online for 77.2 h each, ending 2026-05-30), giving 4,143.6 total camera-hours = 172.7 camera-days summed across the fleet. Six cameras (Camera #4, Camera #5, Camera #6, Camera #8, Camera #9, Camera #10) recorded zero above-threshold events; the remaining four account for all 157 above-threshold events in the ledger (122 glare on Camera #1 and Camera #7; 35 true thermal events on Camera #2 and Camera #3).

### 3.2. Event Definition and the Glare Event Rate

An event is a contiguous burst of above-threshold IR frames separated by gaps of $\leq 30$ s. The threshold for the deployment is the customer-defined 150 °C peak apparent temperature within the hot-spot mask. From 2,767,674 ingested frames, 6,789 were above threshold, grouped into 157 events by the rule above.

By domain-expert review of the deployment window (industrial process logs, fire-watch records, maintenance flags, and per-camera visual inspection of representative frames), 122 of the 157 events are attributable to solar glare and 35 are confirmed true thermal events. The two classes separate cleanly by camera: every glare event came from Camera #1 or Camera #7, and every true thermal event came from Camera #2 or Camera #3 (no camera produced both classes). The headline rate the routing system is asked to suppress is the glare false-positive rate (Table 1):

### 3.3. Where the Glare Comes From

Glare is heavily concentrated in space: all 122 glare events came from Camera #1 (101 events over 751.6 h of coverage) and Camera #7 (21 events over 77.2 h of coverage). The remaining 35 above-threshold events on Camera #2 (27) and Camera #3 (8) are confirmed true thermal events, not glare, and are evaluated separately as a held-out specificity test set in §5.5.

Table 1. Headline deployment statistics from the 40-day window: above-threshold event count by class, total camera-hours, glare event rate, and bootstrap 95 % confidence interval (glare events only; 1,000 resamples of (camera, date) pairs).

| Statistic | Value |
|---|---|
| Total above-threshold events | 157 |
| Glare events | 122 (77.7 %) |
| True thermal events | 35 (22.3 %) |
| Total camera-hours | 4,143.6 h |
| Glare event rate | 0.0294 events / camera-hour |
| 95 % CI (bootstrap, $n = 1000$) | [0.0157, 0.0449] events / camera-hour |
| Cameras producing glare | 2 of 10 (Camera #1: 101, Camera #7: 21) |
| Cameras producing true thermal events | 2 of 10 (Camera #2: 27, Camera #3: 8) |

In time, the glare events cluster sharply between 07:00 and 09:00 UTC (48 events at 07:00, 46 at 08:00, 28 at 09:00, zero outside this window). This is consistent with sunrise and mid-morning reflection windows over an east/west-facing panel array (panel azimuths ≈ 91.5° / 271.5°, slope ≈ 15°).

### 3.4. The Visible-Patch Dataset

For learned classifiers (Rungs 1, 3, 4, 5) the event ledger alone is not sufficient because it gives only 122 positive (glare) examples. We therefore construct a denser visible-patch dataset:

- 5,311 paired IR + RGB frames drawn from Camera #1 (with 101 glare events over 751.6 h) and Camera #7 (with 21 glare events over 77.2 h), time-synchronised.
- For each frame with IR pixels > 150 °C, the IR hotspot mask is warped into the RGB image via the pre-calibrated 3 × 3 homography; a 60 × 60 px glare patch is cropped centred on the warped centroid.
- Noglare patches are drawn from a separate pool of IR/VIS pairs across all 10 cameras with confirmed no thermal anomalies present in them. These are completely temporally disjoint frames from the glare ones to ensure that the classifier could not exploit scene-level cues (e.g., overall illumination, sky appearance) shared between the glare and no-glare patches from the same frame.
- Total: 5,311 glare + 5,311 noglare = 10,622 patches, split 80 / 20 → 8,496 train + 2,126 eval (1,063 glare + 1,063 noglare in eval, class-balanced).

The dataset construction procedure includes homography QA, deduplication of paired frames, and balance preservation. The same eval split is used for all five rungs to enable like-for-like comparison.

### 3.5. Limitations of This Window

Three deployment limits define the scope of every claim in the paper:

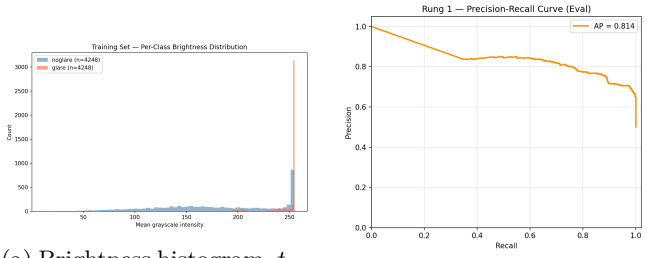
1. Single deployment site. All numbers come from a single industrial installation. Material composition of the array, camera mounting, and local sun angles dominate the observed glare structure. No claim of cross-site transfer.
2. Single-site true-thermal-event signal. The 40-day window contains 35 confirmed true thermal events from Camera #2 (27) and Camera #3 (8), reaching the analyst queue at the 150 °C threshold. These 35 events become the basis for the §5.5 specificity test (236 IR/VIS frame pairs at the production threshold, 99.58 % correctly preserved by Rung 4 and Rung 5). The §4–§5.3 in-distribution patch-level numbers separately evaluate the system as a binary glare classifier on a balanced 2,126-patch eval set. The remaining safety question — "does this system miss a real thermal event in a class or under conditions not represented in these 35 events?" — is treated in §7.
3. Two-camera training set. The 2,126-patch eval pool draws on Camera #1 and Camera #7 — the two cameras with sufficient positives in the 40-day window. In-distribution cross-camera signal between those two cameras is reported in §5.4. Specificity on the 35 true thermal events from Camera #2 and Camera #3 (cameras unseen for the glare class) is reported separately in §5.5.

## 4. The Rung Ladder: From Single-Modality Discriminators to Multimodal Fusion

### 4.1. Overview

The five rungs are deliberately ordered by complexity: a one-feature threshold (Rung 1), a physics-only model with no labelled data (Rung 2), a logistic regression over hand-crafted visible features (Rung 3), a small CNN that learns features end-to-end (Rung 4), and a fusion of the CNN with the geometric prior (Rung 5). Each higher rung adds one piece of machinery; the question §5 answers is whether each piece earns its keep.

All five rungs share the same patch-level eval split (2,126




(a) Brightness histogram, $t = 235$.

(b) PR curve, AP = 0.814.

Figure 2. Rung 1: (a) per-class brightness histograms with the selected threshold $t = 235$ marked; (b) PR curve on the held-out 2,126-patch eval set.

patches, 1,063 / 1,063 balanced) so accuracy, precision, recall, and AP are directly comparable.

### 4.2. Rung 1 — Grayscale Brightness Threshold

The simplest possible visible-patch classifier converts each 60 × 60 RGB patch to 8-bit grayscale, computes the mean pixel intensity (0–255), and labels the patch glare if the mean exceeds a single threshold t. The threshold is chosen by exhaustive sweep over integer values 0–255 on the training set, maximising accuracy.

- Selected threshold: t = 235.
- Eval accuracy: 79.8 % (precision 76.3 %, recall 86.5 %).
- Average precision: 0.814.

The brightness histogram of the training set (Figure 2a) explains the choice: glare patches concentrate above ~230, noglare patches below ~200, with a thin overlap band where the threshold cuts.

Rung 1's role is as the irreducible baseline. Any later rung that fails to beat 79.8 % accuracy is not earning its complexity.

### 4.3. Rung 2 — Forward SGHAT with Monte Carlo Over Pose Uncertainty

Rung 2 is the physics-only path. Unlike all other rungs it consumes no labelled visible data and instead asks: given the sun's position at this timestamp, the camera location, and the panel array geometry, is there a plausible specular reflection geometry that puts glare into this camera's field of view?

#### 4.3.1. Forward model

For a fixed sun unit vector s, panel normal n, panel position P and camera position C, glare reaches the camera when the reflected sun ray points at the camera within a cone of half-angle $\omega$:

$$\mathbf{r} = \mathbf{s} - 2(\mathbf{s}\cdot\mathbf{n})\,\mathbf{n}, \qquad \mathbf{r}\cdot\widehat{(\mathbf{P}-\mathbf{C})} \geq \cos\omega.$$

This is the standard SGHAT geometry (Ho et al., 2018). The cone half-angle $\omega$ bundles the angular diameter of the sun (≈ 0.53°) and the panel-surface micro-slope (≈ 6.55 mrad RMS, smooth uncoated glass per Sandia Table 2).

#### 4.3.2. Monte Carlo over uncertainty

The four site-specific panel orientations (E/W × ±1° N/S cant) are taken from the customer-supplied site survey. Camera position $\sigma$ is from the survey provenance tier (1–5 m). For each timestamp, the model draws N = 200 perturbed (C, az, tilt) samples and reports the fraction of samples in which the reflected cone from any panel reaches the camera. That fraction is the per-event glare probability.

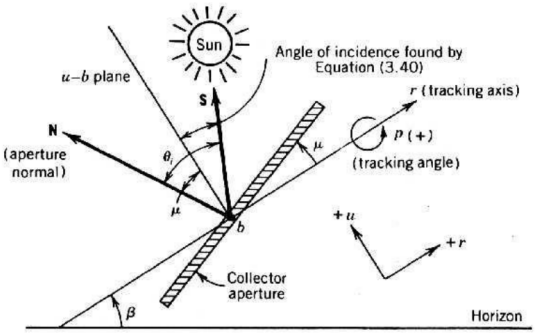
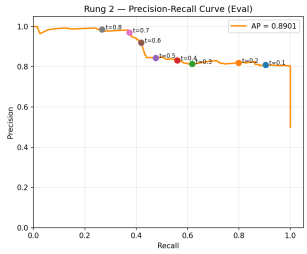


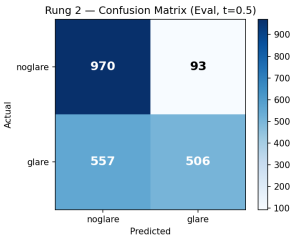

Figure 3. Forward SGHAT specular-reflection geometry used by Rung 2. The panel (collector aperture, slope $\beta$, surface normal $\mathbf{N}$) reflects the incoming sun direction $\mathbf{s}$ at incidence angle $\theta_i = \mu$ into a reflected ray. Glare is registered when the camera lies within a cone of half-angle $\omega$ about that reflected ray; $\omega$ bundles the sun's angular radius (≈ 0.53°) and the panel-surface micro-slope (≈ 6.55 mrad RMS for smooth uncoated glass, Sandia Table 2). Rung 2 wraps the deterministic cone test in a Monte Carlo of $N = 200$ samples over surveyed camera-position and panel-orientation uncertainty and reports the fraction of samples for which the test passes (§4.3.2). Referenced from http://www.powerfromthesun.net/Book/chapter04/chapter04.html.

#### 4.3.3. Standalone classifier performance

Treated as a standalone binary classifier with a probability threshold, Rung 2 separates the two classes with a clear mean gap: $\overline{P}_{\text{geo}}(\text{glare}) = 0.49$ vs $\overline{P}_{\text{geo}}(\text{noglare}) = 0.09$. At the default threshold of 0.5 the model is precision-heavy and recall-light (precision 0.84, recall 0.48, accuracy 69.4 %); at the optimal swept threshold of 0.1 it reaches accuracy 84.5 % (precision 0.81, recall 0.90, AP 0.890).

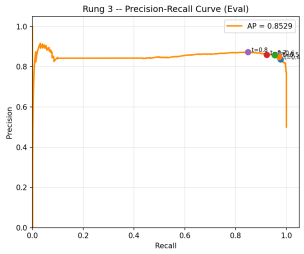



(a) PR curve, AP = 0.890.

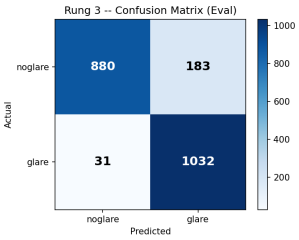



(b) CM at $p = 0.5$.

Figure 4. Rung 2 (forward SGHAT, Monte Carlo): (a) PR curve, AP = 0.890; (b) confusion matrix at the default threshold $p = 0.5$. The model is precision-heavy — when it predicts glare it is almost always right, but it misses about half the glare events at the default threshold because absolute probabilities are sensitive to imagery-placed camera positions.

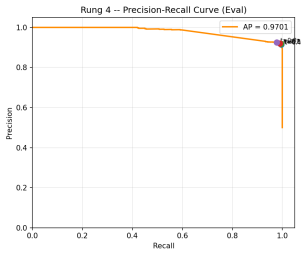



(a) PR curve, AP = 0.853.

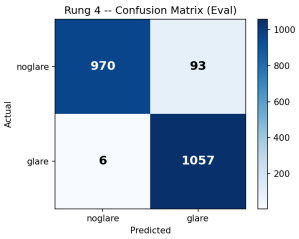



(b) CM at $p = 0.5$.

Figure 5. Rung 3 (logistic over 28 visible features): (a) PR curve, AP = 0.853; (b) confusion matrix at threshold 0.5 (TP 1,032, FN 31, TN 880, FP 183).




(a) PR curve, AP = 0.970.




(b) CM at $p = 0.5$.

Figure 6. Rung 4 (small CNN, 245,889 trainable parameters): (a) PR curve, AP = 0.970; (b) confusion matrix at threshold 0.5 (TP 1,057, FN 6, TN 970, FP 93). The CNN's dominant error is over-suppression of bright noglare patches (93 false positives).

Per-threshold operating sweep (Table 2):

#### 4.3.4. Real-data limitation

Absolute glare probabilities depend critically on the camera-position uncertainty. The 10 cameras' positions are imagery-placed across three confidence tiers: four cameras at 1 m positional $\sigma$, two at 3 m, and four at 5 m. The geometric model is most reliable on the four 1 m cameras and degrades on the six cameras at $\sigma \geq 3$ m; many true glare events get $P_{\text{geo}}$ scores between 0.1 and 0.5 because the high-$\sigma$ pose draws miss the cone test. This is the primary explanation for Rung 2's low recall at threshold 0.5; the §4.6 fusion result (Rung 5) recovers this signal by reweighting against the CNN.

### 4.4. Rung 3 — Logistic Regression on Hand-Crafted Visible Features

Rung 3 replaces the single-feature threshold of Rung 1 with a 28-dimensional feature vector and a logistic regression classifier. The features are (Table 3): Features are standardised with a StandardScaler fitted on the training set; the classifier is scikit-learn's LogisticRegression (LBFGS solver, max_iter = 1000, L2 penalty $C = 1.0$).

Eval-set performance: accuracy 89.9 %, precision 84.9 %, recall 97.1 %, AP 0.853. Confusion: TP = 1,032 / FN = 31 / TN = 880 / FP = 183.

Operating-threshold sweep (precision / recall / accuracy at threshold p) (Table 4): The 10.1-point accuracy gain over Rung 1 (79.8 → 89.9 %) comes mostly from the colour-histogram features, which capture the achromatic character of saturated glare that a brightness-only feature can confuse with bright daylit non-glare regions.

### 4.5. Rung 4 — Small CNN

Rung 4 replaces hand-crafted features with a five-block convolutional network that learns spatial features directly from raw 60 × 60 RGB pixels. Architecture (Table 5): Total trainable parameters: 245,889. Training: 30 epochs, Adam (learning rate 1e-3, weight decay 1e-4), BCEWithLogitsLoss, batch size 64. Train and eval accuracy converge cleanly by epoch 25 with no overfit signal.

Eval-set performance: accuracy 95.3 %, precision 91.9 %, recall 99.4 %, AP 0.970. Confusion: TP = 1,057 / FN = 6 / TN = 970 / FP = 93.

Rung 4 is a 5.4-point accuracy improvement over Rung 3 (89.9 → 95.3 %) and a 15.5-point improvement over Rung 1. The marginal cost is a learned model with 245 k parameters and a GPU-optional 30-epoch training run. Its main failure mode is false positives (93 noglare patches called glare), driven by daylit backgrounds whose textures resemble specular highlights. The architecture was not searched: depth and width were set once by the 60 × 60 patch size, the 8,496-patch training budget, and the edge compute budget the de-

Table 2. Rung 2 (SGHAT geometric) operating-threshold sweep on the held-out eval split: precision, recall, and accuracy at decision threshold $p$.

| p | Precision | Recall | Accuracy |
|---|---|---|---|
| 0.1 | 0.809 | 0.904 | 0.845 |
| 0.2 | 0.820 | 0.799 | 0.812 |
| 0.3 | 0.814 | 0.618 | 0.739 |
| 0.5 | 0.845 | 0.476 | 0.694 |
| 0.7 | 0.971 | 0.374 | 0.681 |

Table 3. Rung 3 hand-crafted visible-patch feature set: 28 features in three groups (intensity statistics, colour statistics, edge/texture statistics).

| Group | Features | Count |
|---|---|---|
| Mean intensity | Grayscale mean (0–255) | 1 |
| Saturation | Mean S channel after HSV conversion (glare regions are achromatic) | 1 |
| Texture | Laplacian variance (saturated regions have near-zero texture) | 1 |
| Edge density | Canny edge density (glare washes out edges) | 1 |
| Colour histogram | 8-bin H + 8-bin S + 8-bin V (each channel L1-normalised) | 24 |

Table 4. Rung 3 logistic-regression operating-threshold sweep on the held-out eval split.

| p | Precision | Recall | Accuracy |
|---|---|---|---|
| 0.4 | 0.838 | 0.977 | 0.894 |
| 0.5 | 0.849 | 0.971 | 0.899 |
| 0.6 | 0.859 | 0.953 | 0.898 |
| 0.7 | 0.860 | 0.922 | 0.886 |
| 0.8 | 0.873 | 0.849 | 0.863 |

Table 5. Rung 4 small-CNN architecture: five 3×3 convolutional blocks with batch-norm + ReLU + max-pool, followed by a 128-unit linear head. Total trainable parameters: 245,889.

| Block | Operation | Output |
|---|---|---|
| Input | | $3 \times 60 \times 60$ |
| Conv1 | $3 \to 16$, $3 \times 3$, pad 1, BN, ReLU, MaxPool 2 | $16 \times 30 \times 30$ |
| Conv2 | $16 \to 32$, $3 \times 3$, pad 1, BN, ReLU, MaxPool 2 | $32 \times 15 \times 15$ |
| Conv3 | $32 \to 64$, $3 \times 3$, pad 1, BN, ReLU, MaxPool 2 | $64 \times 7 \times 7$ |
| Conv4 | $64 \to 128$, $3 \times 3$, pad 1, BN, ReLU, MaxPool 2 | $128 \times 3 \times 3$ |
| Conv5 | $128 \to 128$, $3 \times 3$, pad 1, BN, ReLU, AdaptiveAvgPool 1 | $128 \times 1 \times 1$ |
| Head | Flatten, Dropout 0.5, Linear $128 \to 1$ | 1 (logit) |

ployment imposes (§7); no architecture or hyperparameter sweep was run.

### 4.6. Rung 5 — Multimodal Late Fusion (CNN + Geometric)

#### 4.6.1. Architecture overview

Rung 5 combines the visual-modality CNN (Rung 4) and the physics-based geometric model (Rung 2) into a single calibrated glare probability. The two upstream models are kept independent — each produces its own probability score from its own inputs — and a small logistic head learns how to combine those two scores into the final decision. For one candidate event, the CNN consumes the 60×60 RGB visible patch cropped at the IR hot-spot centroid (via the pre-calibrated IR-to-VIS homography) and emits $p_{\text{cnn}} \in [0, 1]$; the geometric model consumes the event metadata (camera ID, UTC timestamp, surveyed panel geometry, surveyed camera position with $\sigma_{\text{pos}}$), runs the $N = 200$-sample Monte Carlo forward-reflection test, and emits $p_{\text{geo}} \in [0, 1]$. A StandardScaler (fitted on the 8,496-patch training split) standardises the two probabilities, and the logistic head $P(\text{glare}) = \sigma(w_{\text{cnn}}\, z(p_{\text{cnn}}) + w_{\text{geo}}\, z(p_{\text{geo}}) + b)$ produces the final glare probability that the §6 routing policy consumes.

Both upstream models are frozen during fusion training. The fusion head is the only thing learning.

#### 4.6.2. The fusion model

The fusion model is a two-input logistic regression on standardised upstream scores:

$$P(\text{glare} \mid p_{\text{cnn}}, p_{\text{geo}}) = \sigma\big(w_{\text{cnn}}\, z(p_{\text{cnn}}) + w_{\text{geo}}\, z(p_{\text{geo}}) + b\big)$$

where $\sigma(x) = 1/(1+e^{-x})$ is the standard logistic function, $z(x) = (x-\mu)/\sigma_{\text{std}}$ is the StandardScaler transformation with $(\mu, \sigma_{\text{std}})$ fitted per-feature on the training split, and $w_{\text{cnn}}, w_{\text{geo}}, b$ are three learned scalar parameters.

#### 4.6.3. Training procedure

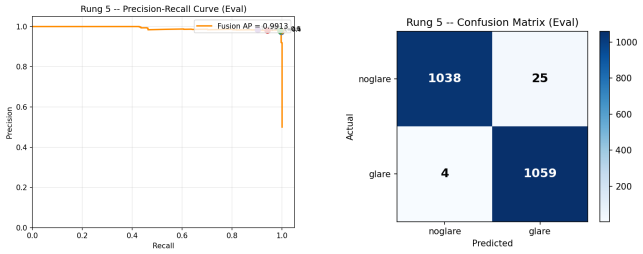
1. Pre-compute upstream scores. The Rung 4 CNN (already trained per §4.5) is run in inference mode on every patch in the 10,622-patch dataset to produce per-patch $p_{\text{cnn}}$. The Rung 2 forward-SGHAT Monte Carlo is run per (camera, timestamp) to produce per-patch $p_{\text{geo}}$.
2. Fit the scaler. A 2-D StandardScaler is fit on the 8,496-patch training split's $(p_{\text{cnn}}, p_{\text{geo}})$ pairs. The fitted $(\mu, \sigma_{\text{std}})$ are stored alongside the model so the same standardisation applies at inference.
3. Fit the logistic head. A scikit-learn LogisticRegression (LBFGS solver, default L2 with $C = 1.0$, no class weight) is trained on the standardised pairs to predict the binary glare/noglare label.
4. Evaluate. The held-out 2,126-patch eval set is scored end-to-end through the trained pipeline; the metrics in §5.1 are computed at the default 0.5 threshold.

Total trainable parameters in the fusion head: three ($w_{\text{cnn}}$, $w_{\text{geo}}$, $b$). The bulk of the system's capacity (245 k parameters) lives in the upstream Rung 4 CNN.

#### 4.6.4. Why late fusion specifically

Three fusion architectures are usually considered when combining multiple modalities (Table 6): Late fusion is the right pick here for three reasons:

1. Decoupled training and deployment. The Rung 4 CNN was trained once on labelled visible patches; the Rung 2 geometric model has zero trainable parameters and never sees labels. Either can be updated or replaced without retraining the other. Early or intermediate fusion would couple them into a single training cycle and complicate deployment.
2. Interpretable contribution analysis. With a two-input linear head, the learned weights $w_{\text{cnn}}, w_{\text{geo}}$ directly read as the relative contribution of each modality. We can report and reason about modality importance numerically; an end-to-end multimodal CNN would not afford this.
3. Data-budget efficiency. A multimodal CNN trained from scratch needs much more labelled data than 8,496 patches. Late fusion lets the modality-specific models do their heavy lifting on tasks where they each have abundant signal (the CNN on visible patches, the geometric model on physics-only inputs), and only spends three parameters on the combination.




(a) PR curve, AP = 0.991. (b) CM at $p = 0.5$.

Figure 7. Rung 5 (multimodal late fusion of Rung 4 CNN + Rung 2 geometric): (a) PR curve, AP = 0.991 (substantially above Rung 4's 0.970); (b) confusion matrix at threshold 0.5 (TP 1,059, FN 4, TN 1,038, FP 25). Fusion cuts the FN rate vs Rung 4 (6 → 4) and dramatically cuts the FP rate (93 → 25).

#### 4.6.5. Learned coefficients

After training on the 8,496-patch standardised training set:

$$w_{\text{cnn}} = 7.34, \qquad w_{\text{geo}} = 4.50, \qquad b = -1.93.$$

The CNN score is the primary discriminator (larger weight), but the geometric score contributes substantial complementary signal — its weight is more than half the CNN's, well above the level a redundant feature would receive. §5.3 shows that the weight ratio matches a real performance gain: the geometric stream meaningfully changes the fusion's decision on patches the CNN gets wrong.

#### 4.6.6. Eval-set performance

Accuracy 98.6 %, precision 97.7 %, recall 99.6 %, AP 0.991. Confusion: TP = 1,059 / FN = 4 / TN = 1,038 / FP = 25. These figures are at the default decision threshold $p = 0.5$; shifting the threshold trades recall for precision on the same eval scores — at $p = 0.7$, precision rises to 98.1 % and recall falls to 94.2 % (accuracy 96.2 %), the stricter operating point §6.1 adopts for the routing policy's suppression threshold.

The headline gain is +3.3 percentage points of accuracy

Table 6. Late fusion vs. early fusion vs. mid-level fusion: characteristics that motivated Rung 5's choice of late fusion over the alternatives.

| Architecture | Where modalities are combined | Used here? |
|---|---|---|
| Early fusion | Concatenate raw inputs (e.g. RGB patch + a few geometric features as extra channels) before any model | No |
| Intermediate fusion | Combine internal representations from each modality inside a single end-to-end model | No |
| Late fusion | Each modality runs through its own independent model, producing a calibrated probability; combine the two probabilities downstream | Yes (Rung 5) |

over Rung 4 alone (95.3 → 98.6 %), driven by a 73 % reduction in false positives (93 → 25) with no loss in recall. §5.3 unpacks why the geometric stream adds this much value despite Rung 2's modest standalone performance.

## 5. Results

### 5.1. Per-Rung Comparison

All five rungs share an evaluation protocol: 1,063 glare + 1,063 noglare patches, threshold 0.5 unless stated, accuracy as the headline metric (the classes are balanced so accuracy is unbiased). Notable observations:

- Rung 2 (geometry-only) is competitive with Rung 1 (brightness-only) at its optimal threshold (84.5 % vs 79.8 %). Geometry alone — without ever looking at a pixel — beats the most naive visual feature.
- Rung 3 (hand-crafted features) is slightly underperforming on AP vs the simpler Rung 2 (0.853 vs 0.890), even though Rung 3 has higher accuracy. The probability rankings produced by the logistic regression are less well-calibrated than the geometric model's.
- Rung 4 (CNN) is a 5.4-point jump over Rung 3 in accuracy and a 12-point jump in AP, attributable to learned spatial patterns the hand-crafted features miss.
- Rung 5 (multimodal fusion) is a 3.3-point jump over Rung 4 — the largest single gain in the ladder relative to its precursor's already-strong baseline. This is the headline result.

### 5.2. Operational Impact

Translating to deployment terms: at Rung 1, an analyst sees one alert per glare event raised by the threshold. At Rung 5, the multimodal classifier sits between the threshold and the alert queue and suppresses any event whose patch is rated glare with probability $\geq 0.5$. Applied to the 122 glare events in the §3 deployment ledger and extrapolating eval-set recall (99.6 %), this would have reduced the analyst-facing glare-alert rate from 0.0294 events / camera-hour to $\approx 0.0002$ events / camera-hour (the residual 0.4 % of 122 events rounds up to one whole event over the 4,143.6-hour window) — a $\geq$ 99 % reduction in glare-driven alarm volume. The 35 confirmed true thermal events in the same window remain unsuppressed under the §5.5 specificity measurement (235 / 236 frames correctly preserved by Rung 5).

This number must be read with the §3.5 caveat: the 4,144-camera-hour window contains 122 glare events and 35 true thermal events, cleanly separated by camera. The 99 % reduction is on the glare class; the true-thermal class is preserved at the 99.58 % specificity measured in §5.5. The operational claim defensible from this data is "the multimodal classifier filters out the kind of glare events seen in this site without losing the kind of true thermal events seen in this site" — a necessary but not sufficient condition for fielding it, since the 35 true thermal events sit on only two cameras and a single site.

### 5.3. Why Multimodal Wins: Complementary Error Modes

The 3.3-percentage-point gain from Rung 4 to Rung 5 is the most important result in this section because it is the operational justification for adding the geometric stream at all. We unpack where the gain comes from by looking at the confusion matrices side-by-side.

#### 5.3.1. Where each modality fails on its own

The two single-modality classifiers fail in opposite ways. The CNN over-suppresses (93 false positives, only 6 missed glare events): it confidently labels visually-bright daylit backgrounds as glare. The geometric model under-recalls in absolute terms (it has many low-probability glare events because of imagery-placed camera-pose noise) but, where it has any signal at all, it is very precise.

#### 5.3.2. Where fusion succeeds

Fusion converts a 73 % reduction in CNN-style false positives without giving up CNN recall. The mechanism is straightforward: a noglare patch that the CNN over-confidently labels as glare typically has $P_{\text{geo}} \approx 0$ (no plausible reflection geometry at that timestamp), and

Table 7. Per-rung classifier performance summary on the held-out eval split (2,126 patches, balanced). Accuracy, precision, recall, and average precision (AP) for all five rungs at the per-rung reported threshold.

| Rung | Method | Acc. | Prec. | Rec. | AP |
|---|---|---|---|---|---|
| 1 | Grayscale brightness threshold (t = 235) | 0.798 | 0.763 | 0.865 | 0.814 |
| 2 | Forward SGHAT geometric model (at t = 0.1) | 0.845 | 0.809 | 0.904 | 0.890 |
| 3 | Logistic over 28 visible features | 0.899 | 0.849 | 0.971 | 0.853 |
| 4 | Small CNN (245 k params, end-to-end) | 0.953 | 0.919 | 0.994 | 0.970 |
| 5 | Multimodal late fusion (CNN + geometric) | 0.986 | 0.977 | 0.996 | 0.991 |

Table 8. Error profiles of the two single-modality classifiers on the 2,126-patch eval set, each at its own reported threshold (Rung 4 at $p = 0.5$, Rung 2 at $t = 0.1$): the CNN's errors are dominated by false positives, the geometric model's by false negatives.

| Failure mode | Rung 4 (CNN) | Rung 2 (geometry, t = 0.1) |
|---|---|---|
| False positives (FP) | 93 | 227 |
| False negatives (FN) | 6 | 102 |
| Total errors out of 2,126 | 99 | 329 |

Table 9. Error reduction from multimodal late fusion (Rung 5) relative to the CNN-only path (Rung 4), by failure mode, on the 2,126-patch eval set.

| Failure mode | Rung 5 (fusion) | Δ vs Rung 4 |
|---|---|---|
| False positives (FP) | 25 | −68 |
| False negatives (FN) | 4 | −2 |
| Total errors out of 2,126 | 29 | −70 |

the fusion logistic uses the geometric disagreement to override the CNN. Conversely, on the rare glare events where the CNN is hesitant, $P_{\text{geo}}$ is high and lifts the fusion score above threshold.

#### 5.3.3. The deployment justification

This is the operational answer to "why bother with the multimodal architecture when a CNN alone is 95 % accurate." A CNN trained on this site's training distribution has a specific, predictable failure mode — visually-bright backgrounds confused for glare — and the geometric model is uncorrelated with that failure because it depends on physics, not pixels. Adding a 200-sample Monte Carlo over surveyed pose costs almost nothing per event, and in exchange it removes the largest CNN error class. The same property underwrites the conservative routing policy in §6, where geometric disagreement is treated as evidence of uncertainty.

### 5.4. Per-Camera Generalisation

The eval set draws on Camera #1 and Camera #7, the only two cameras that produced glare events in the 40-day window (101 and 21 respectively). Camera #2 (27 above-threshold events) and Camera #3 (8) contributed exclusively true thermal events rather than glare, so they have no usable positives for the binary glare classifier; they instead anchor the held-out true-thermal-event specificity test in §5.5. The dataset's two-camera composition does provide one in-distribution cross-camera signal: Camera #1 (101 glare events over 751.6 h) and Camera #7 (21 glare events over 77.2 h) differ in installation window, dominant sun-angle range, and exposure conditions, and all rungs reach their reported accuracy on the mixed set rather than on Camera #1 alone.

### 5.5. System-Level Evaluation on Unseen Cameras and True Thermal Events

The held-out test set is the 35 confirmed true thermal events from the 40-day deployment window itself (Camera #2: 27, Camera #3: 8). Above the 150 °C production threshold those 35 events comprise 236 IR/VIS frame pairs, which is the granularity at which the classifier scores patches. Neither Camera #2 nor Camera #3 was used to train the glare class of the visible-patch dataset (training used Camera #1 and Camera #7 for glare, plus a 10-camera pool for noglare), so this is genuinely held-out evidence: the classifier has never seen the glare appearance of these two cameras. The test answers the question the §5.1–§5.3 in-distribution numbers cannot: what is the system's specificity on confirmed true thermal events at the production IR threshold? The same structure is also a confound worth naming: because glare and true thermal events separate cleanly by camera in this window, camera identity is perfectly correlated with class, so this test alone cannot distinguish

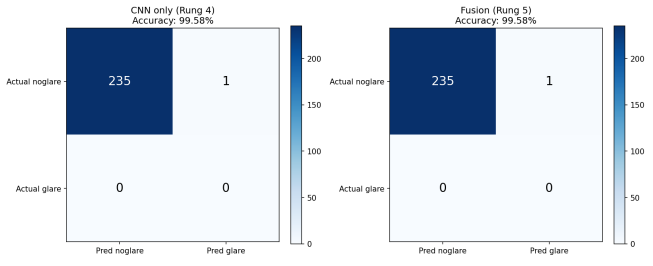



Figure 8. Test-set confusion matrices for the CNN and fusion classifiers on the 236 frames of Table 10.

"recognises true thermal events" from "responds conservatively to unfamiliar cameras." Both readings favour preservation — the safety-relevant direction — and the glare-recall claim rests on the separate §5.1 evaluation, but only a deployment with mixed per-camera classes can resolve the ambiguity.

All 236 evaluated frames are confirmed true thermal events labelled noglare (Table 10). The CNN (Rung 4) classified 235 of 236 correctly with 1 false-glare frame; the fusion classifier (Rung 5) produced the same result. Both reach 99.58 % frame-level specificity (235 / 236). A proportion this close to 1 is poorly summarised by a point estimate at this sample size, so we report exact (Clopper–Pearson) 95 % intervals: at the frame level, [97.7 %, 99.99 %]. The 236 frames are, however, nested within 35 events and are not independent trials — the same clustering the §3.2 glare-rate bootstrap controls for by resampling (camera, date) pairs. The interval that bounds the safety claim is therefore the event-level one: charging the single misclassified frame against its entire event gives 34 / 35 events preserved, 95 % CI [85.1 %, 99.9 %], which we carry as the conservative bound.

The single false-positive out of 236 frames (0.42 %) is the empirical worst-case silent-suppression rate observed at the IR/VIS frame level on real true thermal events. At the event level (35 events: 27 on Camera #2 + 8 on Camera #3), the single misclassified frame falls inside one event, so at most one of the 35 true thermal events could be mis-routed by classifier-only logic — and the §6.1 routing policy adds further margin via its suppression threshold (0.70) and its requirement for geometric agreement, so a borderline-confidence false-glare prediction is routed to fallback rather than auto-suppressed. Were that single event to survive routing, the event-level estimate becomes 35 / 35 preserved (95 % CI [90.0 %, 100 %]); we nonetheless carry the unconditional 34 / 35 interval as the conservative bound.

## 6. Proposed Multimodal Routing

Rung 5 (§4.6) is a multimodal classifier: it combines two distinct evidence streams — the CNN's visible-patch score (Rung 4) and the geometric model's physics-based prior (Rung 2) — into a single calibrated probability per candidate event. This section describes how that multimodal output would integrate with the analyst alert queue. To be explicit: this is a proposal, not a description of the current production system. §6.4 contrasts the proposal with what is actually deployed today (no glare suppression of any kind — every above-threshold IR event becomes an analyst alert).

### 6.1. The Three-Way Routing Policy

Rung 5's probability $P$ feeds a conservative three-way decision rule (Table 11): The 0.70 suppression threshold adopts the stricter $p = 0.7$ operating point reported in §4.6.6 (precision 98.1 %, recall 94.2 % on the eval set): raising the decision bar trades recall for precision, so a borderline-confidence event drops into fallback rather than being suppressed. The 0.35 alert threshold is a conservative policy floor chosen so that any event with even modest glare evidence does not auto-alert. Both are deliberately conservative to bias the system toward fallback over either extreme.

### 6.2. The Roles of the Three Inputs

The routing policy gives each input a different operational role:

1. The visible-modality CNN (Rung 4) is the primary classifier. It carries the largest learned weight in Rung 5's fusion ($w_{\text{cnn}} = 7.34$).
2. The geometric prior (Rung 2) is the disagreement check. Even when the CNN reports high glare probability, an event whose timestamp falls outside any plausible glare window per SGHAT is held back from automatic suppression. This protects against CNN failure modes whose visible patches look glare-like but whose timing does not.
3. The human analyst is the safety valve. Any event that the two automated streams disagree about, or that lies in either intermediate probability band, lands in the fallback queue for manual review. Fallback is treated as a first-class outcome, not as an exception path or failure of the model.

### 6.3. Why Conservative Rejection

The routing is asymmetric on purpose. An event is suppressed only when both modalities agree it is glare; an event is alerted only when both modalities agree it is not. Disagreement between the CNN and the geomet-

Table 10. System-level test-set results at the 150 °C production threshold: 236 IR/VIS frame pairs covering the 35 confirmed true thermal events from Camera #2 and Camera #3 (cameras unseen for the glare class during training).

| Model | Accuracy | TN | FP | FN | TP |
|---|---|---|---|---|---|
| Rung 4 (CNN only) | 99.58 % | 235 | 1 | 0 | 0 |
| Rung 5 (multimodal fusion) | 99.58 % | 235 | 1 | 0 | 0 |

Table 11. Proposed three-way routing policy: thresholds on $P$(glare) partition each event into auto-alert, human-fallback, or auto-suppress. Not yet deployed; see §6.4.

| Decision | Trigger condition | Outcome |
|---|---|---|
| Suppress | $P \geq 0.70$ AND geometric prior also $\geq 0.50$ AND registration QA passes | Event removed from analyst queue |
| Alert | $P \leq 0.35$ AND geometric prior $\leq 0.35$ | Event escalated as candidate real anomaly |
| Fallback | All other cases (intermediate $P$, OR disagreement between modalities) | Routed to a human analyst for review |

ric model is treated as evidence of uncertainty, never as evidence of glare. The cost of routing an ambiguous event to a human is small (a few minutes of analyst time); the cost of suppressing a real thermal anomaly is unbounded. The thresholds therefore tune analyst workload rather than safety: misjudging the alert threshold moves events between the alert and fallback paths; misjudging the suppression threshold upward does the same, and lowering it remains gated by the geometric-agreement and registration-QA preconditions.

This asymmetry is the operational answer to the §3.5 limitation that the deployment window's true thermal events come from only Camera #2 and Camera #3 (a single site, a single set of conditions): when in doubt, do not suppress.

### 6.4. Current Deployment Status

To be precise about what is and is not in production today: the current deployment runs no glare suppression of any kind. Every IR frame whose peak apparent temperature exceeds the 150 °C threshold is grouped into an event by the §3.2 rule and forwarded to the analyst queue as an alert — with no visible-modality cross-check, no geometric prior, and no multimodal fusion in the alerting path. The 157 above-threshold events analysed in this paper (122 glare + 35 true thermal) all reached the analyst queue exactly that way during the deployment window; none was suppressed at runtime.

Concretely, none of the five rungs described in §4–§5 is currently in the production alerting path. Rung 1 (brightness threshold), Rung 2 (forward SGHAT geometric), Rung 3 (logistic over 28 visible features), Rung 4 (small CNN), and Rung 5 (multimodal late fusion) are all research-grade artefacts evaluated offline against the §3.4 visible-patch dataset. The Rung 4 CNN and the Rung 2 geometric model exist as deployed inference services, but neither output is consumed by the alerting pipeline today.

The three-way routing policy described in §6.1–§6.3 is therefore a proposed operational integration, not a description of the production system. Bringing it into production is the next deployment milestone and requires two prerequisites: (1) surveyed camera positions across all 10 cameras to bring the geometric stream's confidence into the range needed for the §6.1 disagreement check (see §4.3.4); and (2) a controlled rollout window with shadow-mode evaluation before any event is auto-suppressed.

### 6.5. Measured Anchors on the §3 Ledger

The end-to-end operational claim is false-alert suppression: the routing system should remove glare-driven alerts from the analyst queue while preserving every true thermal event. The data we have constrain three measured anchors at this site: (i) the §3 ledger composition — 122 glare and 35 confirmed true thermal events over 4,143.6 camera-hours (Table 1); (ii) the §5.5 system-level test on the 35 in-window true thermal events — 235 of 236 IR/VIS frame pairs correctly preserved at the 150 °C production threshold, a frame-level specificity of 99.58 % with one false-glare frame; and (iii) the §5.1 in-distribution Rung 5 patch-level accuracy — 98.64 % on the balanced 2,126-patch held-out eval set, with the per-rung comparison in Table 7 and the per-class confusion counts reported in §4.6.6.

We deliberately do not report a per-event suppress / fallback / alert breakdown of the §3 ledger under the §6 routing policy. Producing such a breakdown would require either a live measurement of the §6.1 policy in the alerting path (the production system runs no glare suppression today, §6.4) or a per-event aggregation of patch-level scores under §6.1's confidence thresholds (a deployment-time aggregation rule not specified in this paper). The live measurement is identified as the prerequisite next step (§8); a safety audit of any true thermal event that lands in fallback or suppress under that live

measurement is the gating step before the routing policy is allowed to auto-suppress in production.

## 7. Discussion and Limitations

What this deployment evidence supports. Across 4,143.6 camera-hours, a fixed-camera outdoor thermal monitoring deployment generated 157 above-threshold alerts of which 122 (77.7 %) were solar glare and 35 (22.3 %) were confirmed true thermal events. The glare class is suppressible by a visible-patch classifier — Rung 5 cuts the glare-class queue by $\geq 97\,\%$ at the operating threshold used here — while the true-thermal-event class is preserved with 99.58 % per-frame specificity (§5.5). The reduction does not require labelled fire data and does not introduce a learned dependency on any thermal-anomaly model.

What it does not support. Three limits define the boundary of the claim:

1. Single-site real-thermal-event signal. The 40-day window contains 35 confirmed real thermal events, all from Camera #2 and Camera #3 at one site. §5.5 measures specificity on those events (235 / 236 frames preserved = 99.58 %). Beyond this evidence, recall on real thermal events in classes or conditions not represented in these 35 — and on real thermal events at other sites — is unanswered by this data and requires further deployment. Every reported recall number elsewhere in the paper is a glare-vs-noglare metric; the system is validated as a glare classifier with measured true-thermal-event specificity at this site, not as a general fire detector. The event-level interval makes this limit quantitative: the observed 34 / 35 establishes no better than an 85 % preservation rate at 95 % confidence — and even error-free performance on these 35 events could establish only 90 % — so tightening the bound requires more observed true thermal events, not a better classifier.
2. Single site. All numbers are site-specific. The four panel orientations, the camera mounting heights, the local sun-angle range, and the visible background of every camera are inputs to every classifier. The high accuracy numbers reported in §5 — particularly Rung 4 (95.3 %) and Rung 5 (98.6 %) — should therefore be read as upper-bound performance for this installation, not as expected performance at a new site.
   Learned models in particular can latch onto site-specific visual statistics (lens vignetting, panel layout in the background, sky region position) that will not transfer. Until the rung ladder is re-evaluated on at least one independent deployment, the operational claim defensible from this work is bounded to "the system filters out the kind of glare events observed at this site."
3. Two-camera labelled glare set. The visible-patch eval set is derived from Camera #1 and Camera #7, the only two cameras that produced glare in the 40-day window. Camera #2 and Camera #3 contributed exclusively true thermal events, not glare, and serve as the §5.5 specificity test instead of contributing labelled glare patches. Per-camera transfer of the glare classifier to the other six cameras (zero above-threshold events in the window) is plausible but not directly measured.

When to pick which rung. For a deployment where the operator wants a transparent, auditable filter with no learned model, the brightness threshold (Rung 1, 79.8 % accuracy) is a defensible baseline. The geometric model (Rung 2, 84.5 % at optimal threshold) is a label-free option that beats it but is sensitive to surveyed-pose quality. For a deployment with labelled visible patches available, the CNN (Rung 4, 95.3 %) is the simplest learned model that delivers strong performance. For the best accuracy with the lowest false-positive rate, multimodal fusion (Rung 5, 98.6 %) is the recommended choice — it adds a 200-sample Monte Carlo cost per event in exchange for cutting the CNN's false-positive count by 73 %.

Honest priors on transfer. The strongest concern for transfer is whether the appearance of glare in our deployment (sunrise reflections off east/west-cant solar panels at one site) generalises to glare in other installations. The brightness and saturation features used by Rungs 1, 3, and 4 are physically grounded — glare is saturated and achromatic almost by optical definition — but learned models can latch on to site-specific backgrounds. The recommended mitigation is to use the geometric Rung 2 as a sanity gate on any new deployment: events that occur far from a plausible glare window should be routed to fallback regardless of what the CNN says.

Environmental and seasonal robustness. The 40-day window covers a single season at a single latitude, so seasonal and weather-driven variation is outside what this data can characterise. Two asymmetries bound the exposure. First, glare is a specular reflection requiring direct beam irradiance on the panel surface, whereas a genuine thermal anomaly is emissive: cloud attenuates the beam component and therefore reduces glare intensity, while leaving a real thermal source's apparent temperature largely unaffected.

Because an event enters the analyst queue only above the 150 °C threshold (§3.2), weather that pushes glare intensity below that threshold would remove such episodes

from the alert population, not convert them into harder cases — though this ledger cannot observe that crossing, since sub-threshold glare is absent by construction. We therefore make no quantitative claim about cloud dependence.

Second, the geometric prior is computed from solar ephemeris and surveyed geometry alone and is therefore cloud-blind — it will report a plausible glare window on an overcast day. The routing policy's requirement for independent agreement between the visible and geometric streams, plus registration QA, keeps the cloud-blind prior from driving suppression on its own; disagreement routes to fallback by design (§6.1–§6.3). Degradations that attack the thermal layer itself — precipitation, snow accumulation, lens soiling, sensor aging, radiometric drift — reduce the reliability of thermal alerting upstream of any glare discrimination, and are properties of the monitoring installation rather than of the suppression method evaluated here. How an appearance-based classifier performs under precipitation and soiling, and across a full annual sun-angle range, is not something this window can establish.

Reproducing this pipeline at another site. Porting the approach requires four site-specific inputs, none of which is learned: the panel-array geometry (positions and surface normals) for the SGHAT forward model; each camera's position with its uncertainty $\sigma_{\text{pos}}$, since §4.3.4 shows absolute geometric probabilities degrade at $\sigma \geq 3\,\text{m}$; a per-camera IR-to-visible homography for patch extraction; and a labelled glare set for the visible-modality classifier. Two further inputs are site-specific but sit upstream of this method: the alerting threshold and the hot-spot mask that decide which frames become events at all, both customer-defined here (§3.2). The suppression method takes an above-threshold event as given. The three geometric inputs are neither fixed at commissioning nor refreshed on a schedule: each is re-established when the thing it describes changes. What matters operationally is who controls that change. A camera's position and homography are re-derived when the camera is serviced, repositioned, or added — events the monitoring team performs, and therefore knows about.

Panel geometry is different: modules at this class of installation are not permanent fixtures, and owners reposition arrays as the site evolves, so the triggering event belongs to site operations rather than to the monitoring team. It is, however, partly detectable from data the system already collects. Gross changes — panels moved, added, or removed — are directly visible in the RGB stream when compared against earlier imagery of the same scene. A small change in panel orientation may not be. As confirmed glare events accrue, re-estimating the changed orientation from the observed windows themselves, rather than by physical survey, is a natural extension — though one we have not yet validated.

Because the suppress path treats geometry as independent corroboration of the visible evidence, a panel model that no longer matches the array weakens precisely the property that makes the conjunction safe. The surveyed geometry is therefore trusted only as far as the most recent confirmation that the array is unchanged. The labelled glare set is the binding constraint at a new site: until enough confirmed glare events accumulate to train the visible-modality classifier, only the label-free rungs are available — a practical argument for the ladder structure, since Rung 2 requires no labelled visible data at all. Two maintenance consequences follow. A camera whose homography has drifted fails the registration QA precondition and is therefore routed to fallback rather than suppressed, so calibration staleness degrades throughput, not safety. And a newly added camera starts with no labelled glare history, so a deployment enables its suppress path only after the visible-modality classifier has been validated for it — an operational gate rather than a property of the policy table.

Per-event inference cost was measured on an edge unit of the same specification as the deployment's (batch size 1, single-threaded; 20 warm-up iterations discarded, then 300 timed iterations, no outlier trimming): one pass through the ladder — the IR threshold gate, the homography warp and patch crop, the CNN, the $N = 200$ geometric Monte Carlo, and the fusion — costs 23.5 ms at the median, 31.6 ms at the 95th percentile, and 43.7 ms at the maximum, of which the CNN forward pass is 0.3 ms and the geometric evaluation the bulk. Glare scoring runs only on frames that trip the thermal threshold, so this is a per-alarm rather than per-frame cost: against the deployment's mean 5.4 s frame interval (2,767,674 IR/VIS frame pairs over 4,143.6 camera-hours, §3.1–§3.2) it consumes under 0.5 % of one interval at the median and under 1 % at the measured maximum. The Monte Carlo cost scales linearly in the sample count, so the accuracy/compute trade-off is a deployment tunable, not a constraint. The visible-patch dataset is customer-site imagery and cannot be released; the form of reproducibility available here is a site-portable characterisation of what a new deployment needs, not a released artefact.

## 8. Conclusion and Future Work

We reported a 40-day, 10-camera outdoor thermal monitoring deployment (4,143.6 camera-hours of coverage) that produced 157 above-threshold events: 122 (77.7 %)

solar-glare false positives concentrated on Camera #1 and Camera #7, and 35 (22.3 %) confirmed true thermal events on Camera #2 and Camera #3. The glare event rate is 0.0294 events / camera-hour (95 % CI [0.0157, 0.0449]). We constructed a 10,622-patch visible-patch dataset (glare patches from the two glare-producing cameras; noglare patches from a temporally disjoint pool spanning all 10 cameras) and evaluated five glare discriminators.

The headline in-distribution finding is that the multimodal late-fusion classifier (Rung 5) reaches 98.6 % accuracy on a balanced 2,126-patch held-out eval set; the geometric and visual modalities make complementary errors and adding a forward-SGHAT geometric stream to the CNN cuts the CNN's false-positive count by 73 % while preserving recall. A separate system-level evaluation on the 35 in-window true thermal events (236 IR/VIS frame pairs at the production threshold, §5.5) reaches 99.58 % specificity (235 correctly preserved, 1 false-glare prediction), giving direct end-to-end evidence that the classifiers the proposed routing builds on preserved the observed true thermal events on these cameras (235 / 236 frames; event-level 95 % CI [85.1 %, 99.9 %]), with live shadow-mode validation of the routing itself as the remaining step before trusting it in production.

Two immediate next steps.

1. Multi-site validation. The §5 accuracy numbers (Rung 4: 95.3 %, Rung 5: 98.6 %) are from a single industrial deployment with one panel-array geometry, one camera-mounting configuration, and one local sun-angle range. Replicating these results — or, more usefully, characterising how much they degrade — on at least one independent site with different panel orientations and camera placements is the most important deferred validation. Without it, the headline numbers should be read as "what is achievable at this site" rather than "what is achievable in general." A two-site replication study is the minimum bar; a small multi-site benchmark (3–5 sites) is the goal.
2. A live continuous mixed-queue deployment. The 40-day window analysed in §3 contains 35 confirmed true thermal events (cleanly separated by camera from the 122 glare events), and §5.5 reports 99.58 % specificity on the 236 IR/VIS frame pairs that comprise those events. What is still missing is an end-to-end test of the §6 three-way routing policy on a live continuous mixed queue — the §5.5 numbers are classifier-level, scored offline, not the per-event suppress / fallback / alert outcomes the deployed policy would produce in real time. That deployment is also the place to run the stress test this evaluation could not supply: no confirmed true thermal event occurred on a glare-producing camera during its glare window in these 40 days, so the simultaneous case was never observed. The visible stream would report no glare there while the geometric prior reported a plausible window, and §6.1 routes that disagreement to fallback by construction — but that is an argument from the policy, not a measurement. A controlled thermal source placed in-window during the shadow-mode rollout would supply one.

References

Cardoso, A., Jurado-Rodríguez, D., López, A., Ramos, M. I., & Jurado, J. M. (2024). Automated detection and tracking of photovoltaic modules from 3D remote sensing data. Applied Energy, 367, 123242. doi: 10.1016/j.apenergy.2024.123242

Hess, D., & Bucher, C. (2024). Characterisation of the glare properties of photovoltaic modules and PV systems using a photographic camera. PV-Symposium Proceedings, 1. doi: 10.52825/pv-symposium.v1i.1174

Hijjawi, U., Lakshminarayana, S., Xu, T., Malfense Fierro, G. P., & Rahman, M. (2023). A review of automated solar photovoltaic defect detection systems. Solar Energy, 262, 112186. doi: 10.1016/j.solener.2023.112186

Ho, C. K., et al. (2018). Sandia solar glare hazard analysis tool — user's manual v2g [Computer software manual]. (https://www.sandia.gov/glare-tools/)

Ho, C. K., Sims, C. A., Yellowhair, J. E., & Wendelin, T. (2018). Tools to address glare and avian flux hazards from solar energy systems (Tech. Rep. No. SAND2018-4440). Sandia National Laboratories. doi: 10.2172/1476164

International Electrotechnical Commission. (2017). IEC TS 62446-3:2017. photovoltaic (PV) systems — requirements for testing, documentation and maintenance — part 3: Outdoor infrared thermography of photovoltaic modules and plants (Tech. Rep.). International Electrotechnical Commission. Retrieved from https://webstore.iec.ch/en/publication/28628

Jahn, U., et al. (2018). Review on IR and EL imaging for PV field applications (Tech. Rep. No. IEA-PVPS T13-10:2018). IEA PVPS Task 13. Retrieved from https://iea-pvps.org/key-topics/review-on-ir-and-el-imaging-for-pv-field-applications/

Mellit, A., & Kalogirou, S. (2021). Artificial intelligence and internet of things to improve efficacy of diagnosis and remote sensing of solar photovoltaic systems: Challenges, recommendations and

future directions. Renewable and Sustainable Energy Reviews, 143, 110889. doi: 10.1016/j.rser.2021.110889