Illuminex AI
AIRFIELD INTELLIGENCE

TECHNICAL

# WHITE PAPER

## SEEING WHAT MATTERS:

Evaluating FOD$^{AI}$ Beyond Frame-Level Metrics

**Nikita Shymberg,**
**VP of Engineering**

**FEBRUARY 2026**

# Table of Contents


Executive Summary 3

Introduction: The Unique Challenges of Moving-Platform FOD[AI] 4

Why Frame-Level Metrics Fail 5

A Concrete Example: Model Comparison 5

Video-Centric, Object-Level Evaluation Framework 6

- Full-Video Inference
- Object-Level Attribution
- Aggregation into Tracks
- Track Evaluation

Addressing the Coalescence Problem 7

Pipeline Selection and Trade-Offs 8

Future Improvements: Sensor Data Fusion 9

Conclusion 10

# Executive Summary

Foreign Object Debris (FOD) on airport runways remains one of the most persistent threats to aviation safety. Even small items, such as screws, tire fragments, or pavement pieces, can cause engine damage, tire bursts, or catastrophic failure. Modern FOD detection increasingly relies on AI, but evaluating these systems is far from straightforward when deployed on moving vehicles.



## FOD$^{AI}$



FOD$^{AI}$ is a moving platform that scans the entire runway in real time, capturing continuous video under dynamic conditions, including motion blur, rapidly changing perspectives, and small and transient objects. Standard object detection metrics such as mAP@0.50 or frame-level recall were designed for independent still images and fail to capture the operational goal: reliably detecting every unique hazardous object while minimizing unnecessary alerts for inspectors.

This white paper presents a video-centric, object-level evaluation framework specifically developed for moving-platform FOD systems. By preserving full temporal sequences, attributing detections to individual objects, aggregating them into coherent tracks, and measuring object-level recall alongside alert-based false positive rate (FPR), the framework aligns evaluation with real-world outcomes: comprehensive hazard detection with manageable human review burden.

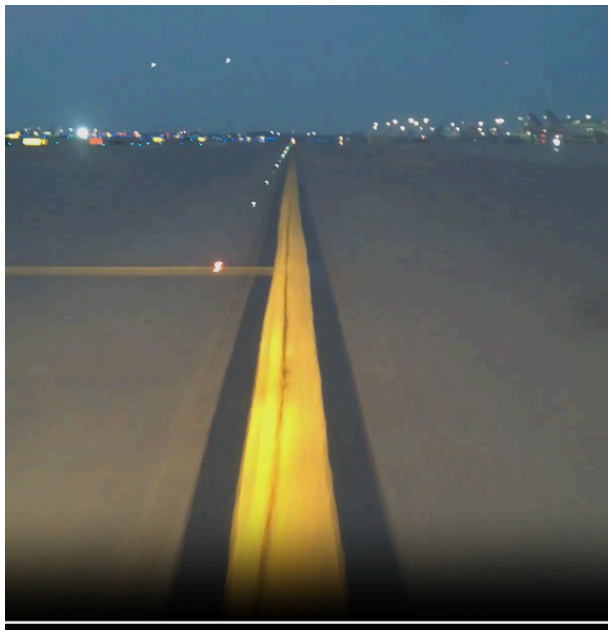
# Introduction

## The Unique Challenges of Moving-Platform FOD[AI]

Airport runway inspections must be fast, thorough, and minimally disruptive. Manual visual checks are labor-intensive, weather-dependent, and prone to human error. Automated systems, especially those mounted on vehicles, offer continuous, high-resolution scanning of entire movement areas.

However, the moving-platform context introduces evaluation challenges not present in fixed-camera or static-image benchmarks:

- Motion blur at typical vehicle speeds (20–40 km/h) can render even large objects indistinct in many frames.
- Small objects, such as bolts, tire fragments, or pebbles, appear only briefly as the vehicle passes.
- Rapid perspective changes cause objects to shift quickly across the field of view.
- High frame-to-frame correlation means the same object is visible across multiple frames, sometimes dozens, while others may appear for only a few.



Not all frames with a FOD label are equal - on one end of the spectrum we have FOD that is so far away from the camera it is indistinguishable from asphalt texture; on the other end, if the FOD is too close to the camera, it is unresolvable due to motion blur. The best frames are somewhere in the middle and conventional frame-level metrics incorrectly treat all frames equally.

To address this disconnect, we developed a video-centric, object-level evaluation framework that:

1. Preserves the complete temporal sequence of predictions.
2. Groups detections into meaningful tracks.
3. Filters noise to reduce spurious alerts.
4. Measures performance at the level of unique objects and human review workload.

# Why Frame-Level Metrics Fail

Standard object detection evaluation treats each frame as an independent image. Precision, recall, and mean Average Precision (mAP) are calculated using intersection-over-union (IoU) thresholds, assuming no temporal relationship between images.
In continuous video from a moving vehicle, this assumption collapses:

- Temporal correlation causes the same object to appear across many frames. A model that locks onto one object and detects it repeatedly scores high on frame-level recall, even if it misses every other object.
- Realistic visibility constraints make perfect per-frame recall impossible. Distant objects are too small to resolve reliably; close objects blur under motion; angled views or partial occlusions further reduce detectability.

A good FOD detector should alert inspectors when an object is clearly visible, not necessarily in every single frame. Frame-level metrics reward redundant detections and penalize realistic intermittent detection, which is the opposite of operational needs.

## A Concrete Example: Model Comparison

Assume a single video contains 5 distinct FOD objects, each visible for 100 frames, giving us a total of 500 labelled bounding boxes.

Now compare two models:

| Model | Behavior | Frame-Level Recall | Unique Objects Detected | Operational Outcome |
|---|---|---|---|---|
| A | Detects only FOD #1 in all 100 frames | 100/500 = 20% | 1 | Misses 80% of hazards |
| B | Detects all 5 objects, 10 consecutive frames each | 50/500 = 10% | 5 | Detects every hazard at least once |

From a frame-level perspective, Model A appears superior. From an operational perspective, Model B is clearly better: it identifies every unique object, enabling inspectors to act on all hazards.

# Video-Centric, Object-Level Evaluation Framework

Evaluating recall one object at a time, we shift from frame-level to video-level, object-level metrics.

Inference is run on every frame of the evaluation videos. The complete time series of predictions is preserved, providing the raw detections for temporal analysis and aggregation.

We have collected a broad evaluation dataset consisting of two kinds of videos: FOD placement videos are short and contain lots manually placed and manually labelled FOD pieces; full inspection videos are much longer and have no FOD in them. The first kind of video helps calculate the model's ability to detect FOD whereas the second kind of video helps estimate the false positive rate

For each labeled FOD instance, detections overlapping any other labeled object are excluded. This enforces clean attribution and prevents one object from "borrowing" detections from nearby objects.

## Aggregation into Tracks

Detections pass through a series of filters including:

- Remove boxes outside the operational region of interest (ROI).
- Remove low-confidence detections.
- Remove isolated, one-off detections that are likely noise.
- Aggregating invidivual detections that are close to each other in time into tracks.

## Track Evaluation

Each track is classified as:

- True Positive (TP): overlaps the labeled FOD instance, even partially.
- False Positive (FP): no match to any labeled FOD.

Key metrics:

- Object-level recall: percentage of unique labeled FOD objects with at least one overlapping track.
- False Positive Rate (FPR): number of FP tracks per unit of video duration.

# Addressing the Coalescence Problem

Coalescence occurs when detections from multiple distinct objects are grouped into a single long track. Highly sensitive pipelines often produce high recall but merge nearby objects, creating long, difficult-to-review clips.

The operational impact of long clips is that a five-minute merged track may contain three separate hazards. Inspectors must scrub through the entire clip to find them, wasting time and risking missed objects.

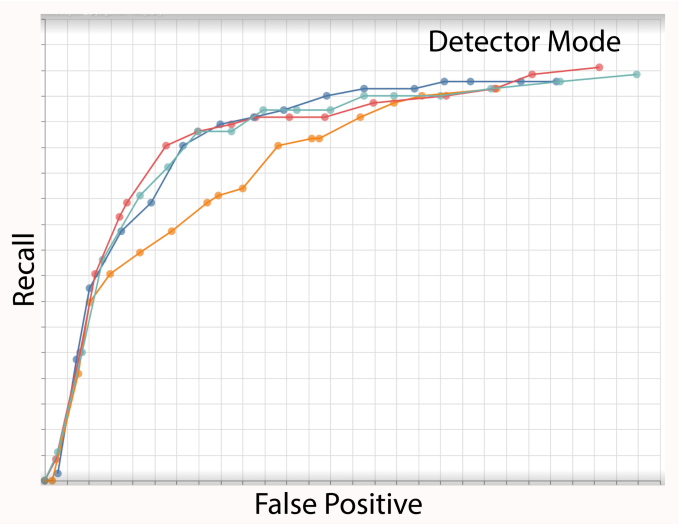
## Why Standard Tracking Fails

There are industry standard object trackers available that can distinguish individual objects from a sequence of detections.

- Small objects combined with fast vehicle motion result in minimal bounding box overlap between frames.
- Limited visual features inside small bounding boxes make appearance-based tracking unreliable.

## Mitigation Strategies

To mitigate this problem, we enforce the maximum expected clip duration to no more than 30 seconds. As a more robust, long-term solution, we will use sensor fusion to assign real-world GPS coordinates to detections and separate them into individual obejcts using clustering algorithms.



# Inference Pipeline Selections and Trade-Offs

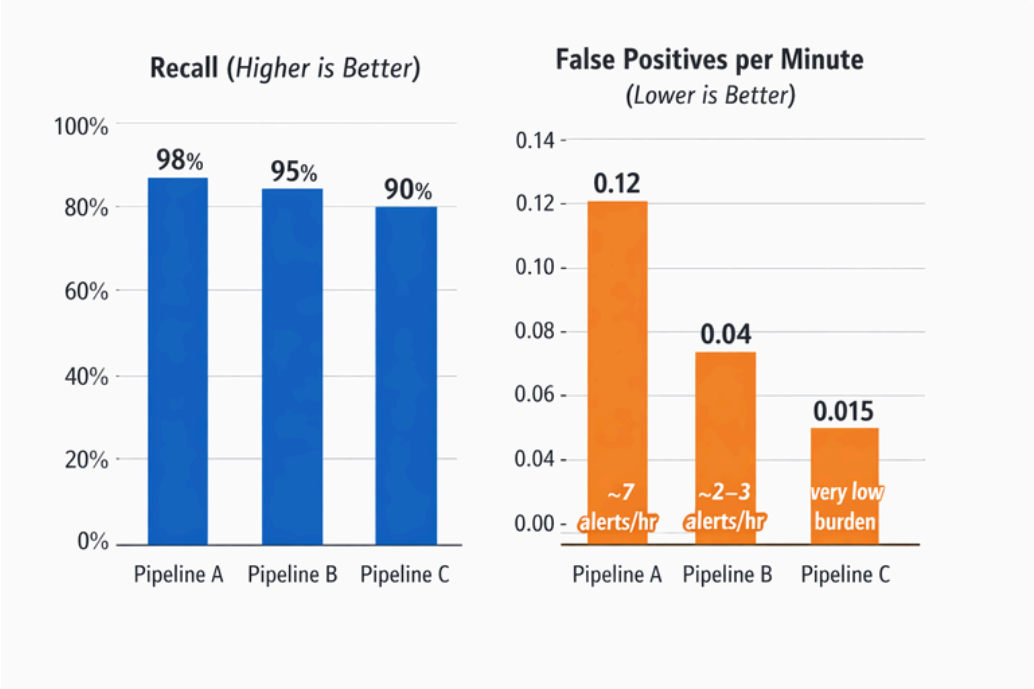
Inference pipeline configuration involves many parameters, including confidence thresholds, temporal proximity windows, and parameters defining the region of interest.

By plotting object-level recall against alert-based FPR in a Pareto chart, teams can visually identify trade-offs:

- High-sensitivity pipelines produce high recall but longer or more tracks, increasing review burden.
- Conservative pipelines reduce recall but generate fewer and shorter tracks, reducing workload.

## Example trade-off points:

- Pipeline A: 98% recall, 0.12 FP tracks per minute (~7 alerts per hour).
- Pipeline B: 95% recall, 0.04 FP tracks per minute (~2–3 alerts per hour).
- Pipeline C: 90% recall, 0.015 FP tracks per minute (very low burden).
- The optimal choice depends on operational priorities: high-traffic airports may favor lower FPR, while lower-traffic sites may accept more alerts for maximum coverage.



# Future Improvements: Sensor Data Fusion

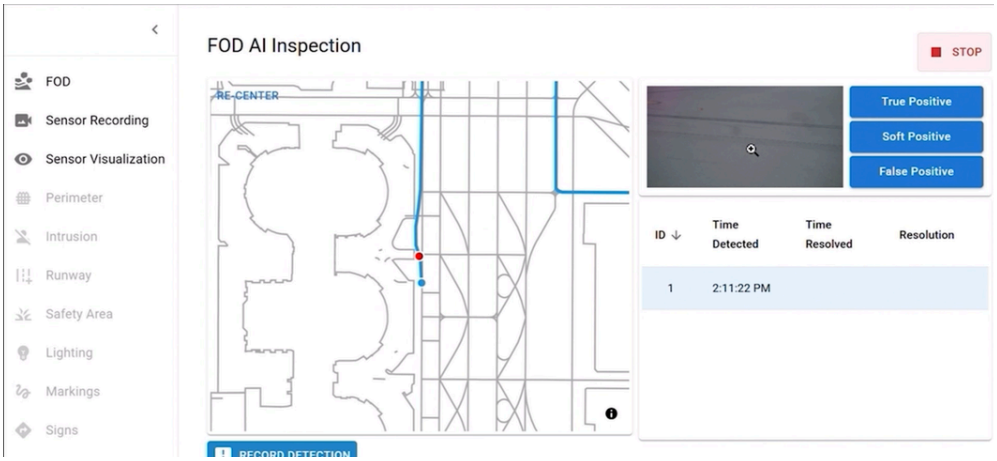
Integrating GPS, IMU, and stereo camera or LiDAR data assigns world coordinates to each detection.

## Key Benefits

- Resolve coalescence by clustering detections by GPS location to separate merged objects.
- Suppress known features by ignoring detections at fixed runway elements, such as lights or signs.
- Increase sensitivity, detecting smaller or more distant objects without inflating FPR.





# Conclusion

Frame-level metrics distort the true performance of moving-platform FOD$^{AI}$ systems. A video-centric, object-level framework, centred on track aggregation, clean attribution, and practical metrics, detecting every unique hazard while minimizing inspector workload.

This approach has already proven effective in beta deployments across multiple airports, delivering hundreds of unique detections with a mix of actionable and non-actionable alerts. By continuing to refine track management and advancing toward sensor fusion, we can build FOD$^{AI}$ systems that are not only accurate but truly production-ready and operationally valuable.

