Self-Guiding Sentinels

Sarah Mundy
srm2238

Objectives

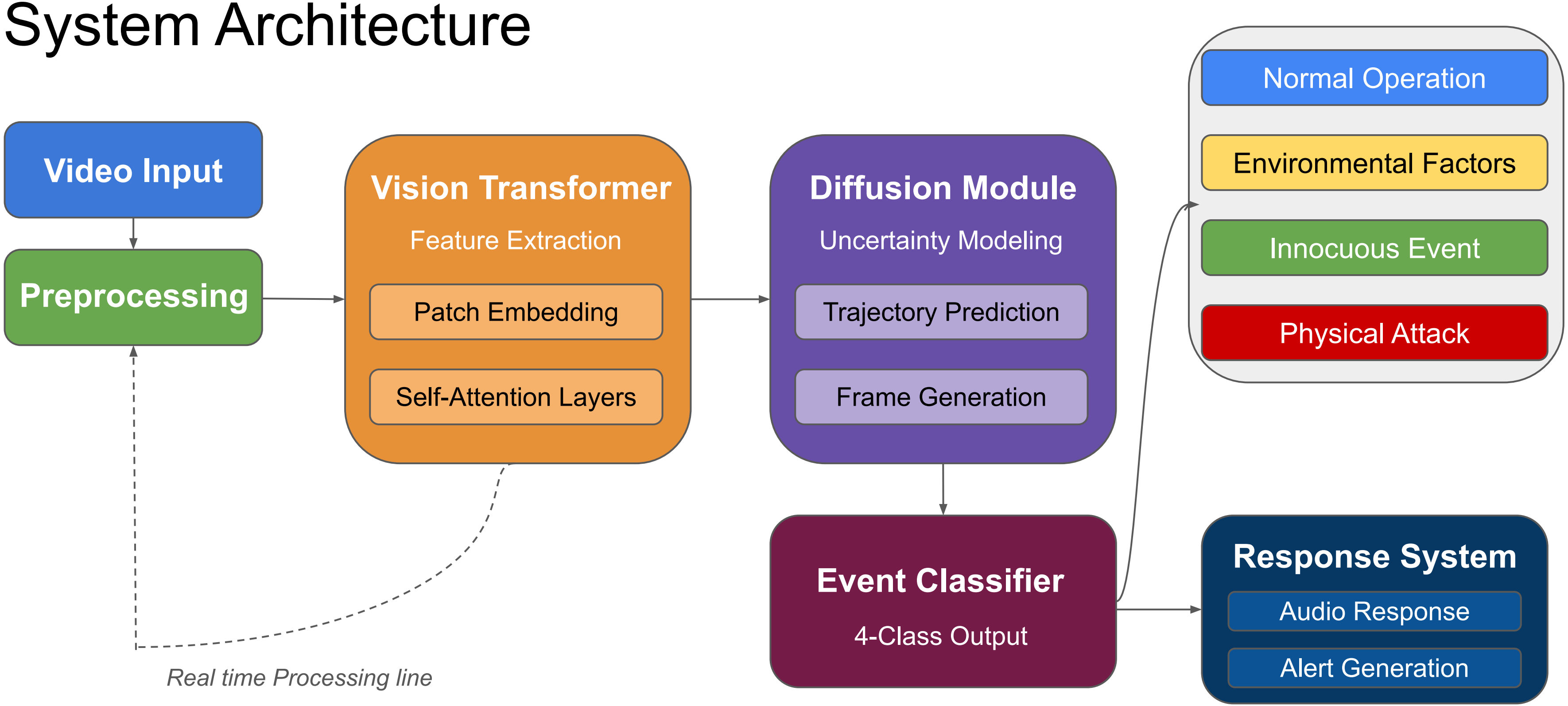
Physical Tampering Attacks

⚠️ Challenge

Physical attacks on NSF streetscape infrastructure require novel detection approaches beyond traditional cyber defences

Real-World Impact:
“Smart trees” destroyed within 3 days of deployment on 125th St in Manhattan October 2024

System Architecture



Response System

System deletes video feed if >10 min old unless marked for review.

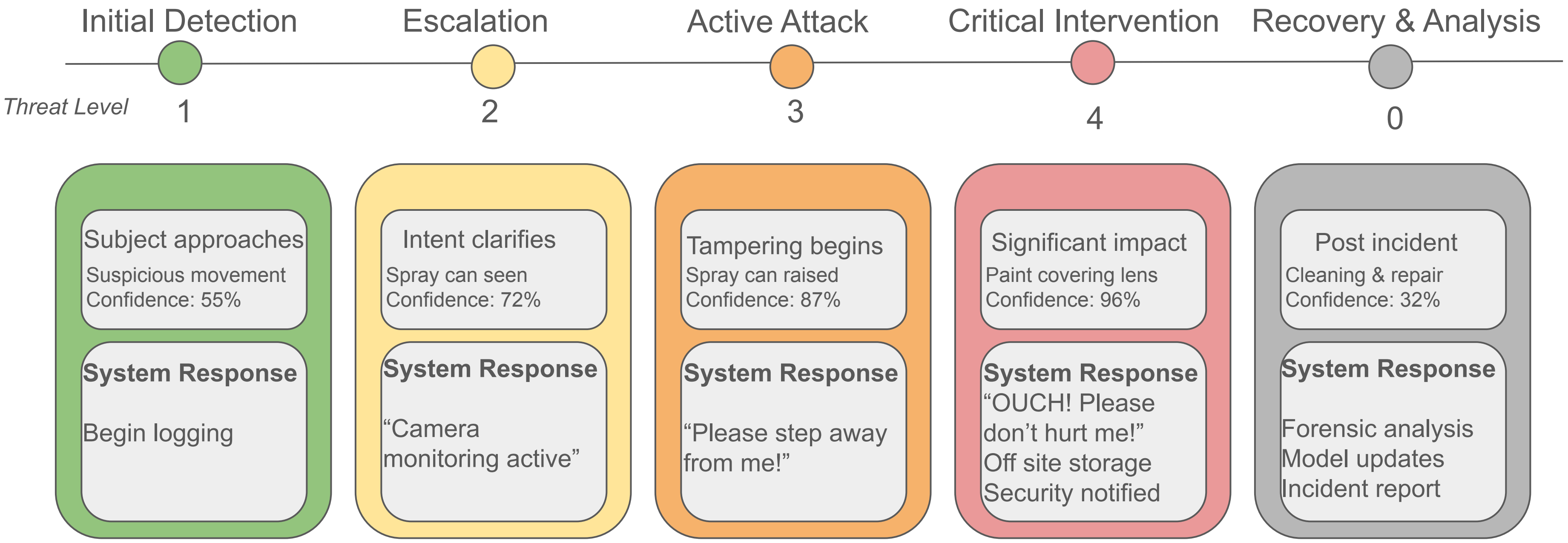
If a threat is noticed, the system saves the last 10 minutes of data for review, retraining, and chain of custody purposes (if the threat was valid)

The maintainer selects who is responsible for receiving and acting on the alerts from the system

Threat Levels

1	**Low Concern** (Monitoring)	Triggers	Low confidence score (50-65%) but potential tampering detected
		Response	Log event with timestamp & confidence score
2	**Moderate Concern** (Warning)	Triggers	Medium confidence tampering detected (65-80%) Multiple low confidence detections within short window
		Response	Soft audio warning: “Camera monitoring active”
3	**High Concern** (Deterrence)	Triggers	High confidence (80-90%) tampering detected
		Response	Medium audio warning: “Please step away from the camera/me”
4	**Critical Threat** (Active Intervention)	Triggers	Very high confidence score (>90%) tampering detected
		Response	Loud audio warning: “Ouch! Please don’t hurt me!” Security personnel notification Save previous 10 minutes of video feed
5	**Breach** (Emergency Protocol)	Triggers	Loss of camera functionality Signal interruption after Level 4 alert
		Response	Request security dispatch

Example Workflow: Physical Attack



Threat Level

Initial Detection

1

Subject approaches

Suspicious movement
Confidence: 55%

System Response

Begin logging

Escalation

2

Intent clarifies

Spray can seen
Confidence: 72%

System Response

“Camera monitoring active”

Active Attack

3

Tampering begins

Spray can raised
Confidence: 87%

System Response

“Please step away from me!”

Critical Intervention

4

Significant impact

Paint covering lens
Confidence: 96%

System Response

“OUCH! Please don’t hurt me!”
Off site storage
Security notified

Recovery & Analysis

0

Post incident

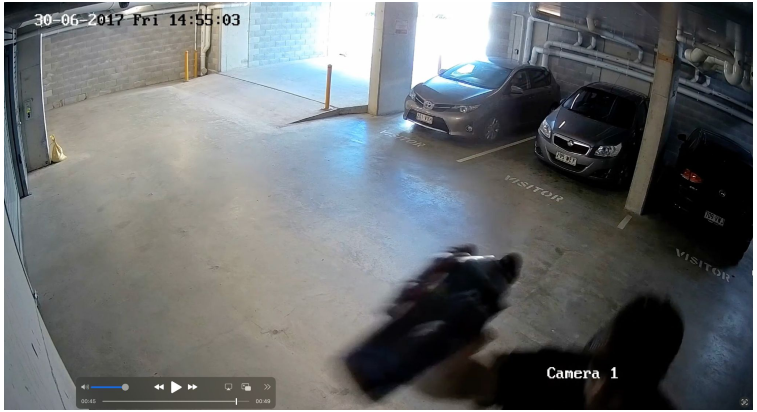
Cleaning & repair
Confidence: 32%

System Response

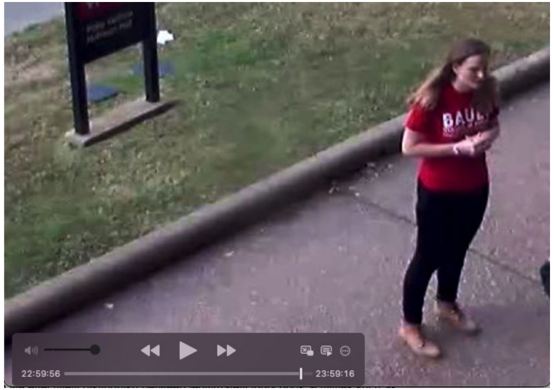
Forensic analysis
Model updates
Incident report

Sample Inputs

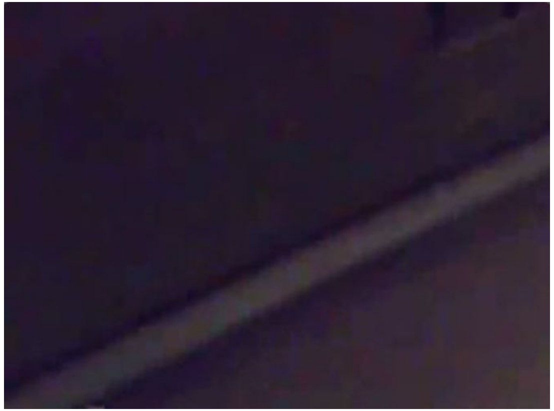
1. Attack



2. Normal Operation



3. Innocuous Event

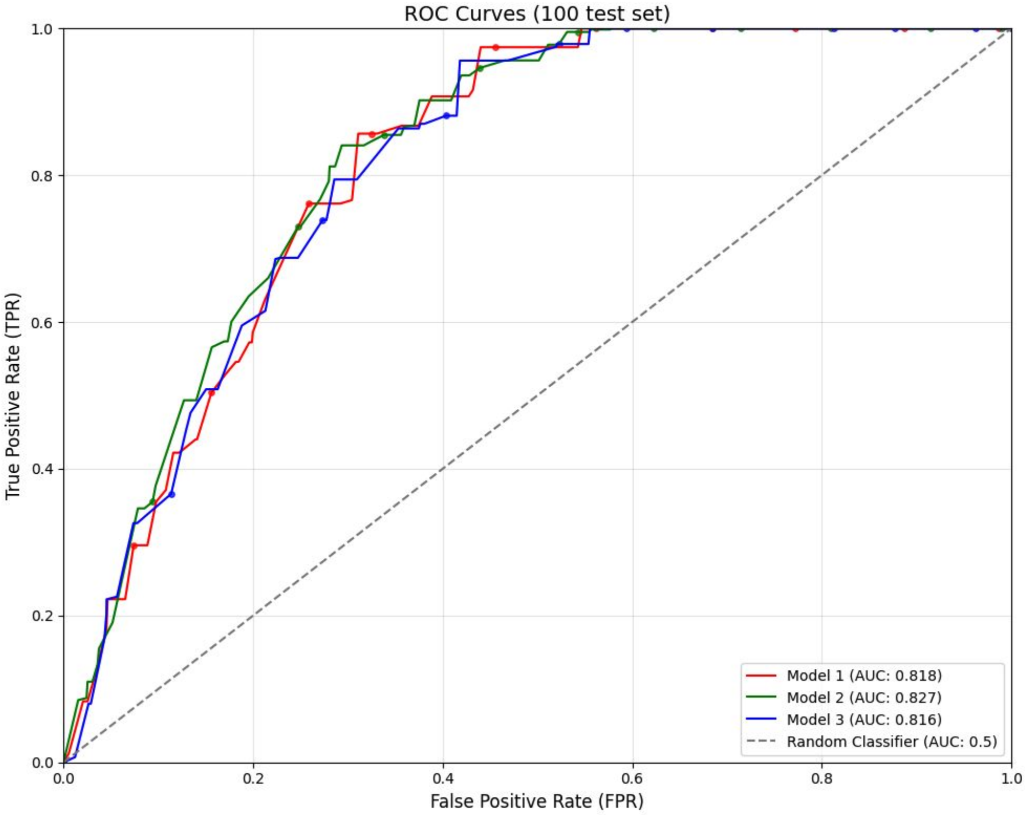


4. Environmental

Outputs

Sample Number	Normal Operation	Environmental Factors	Innocuous Event	Physical Attack
1	0.0241	0.0001	0.4702	0.8957
2	1	0	0	0
3	0.6333	0.6347	0.8587	0.0201
4	0.5861	0.8942	0.6678	0.4431

Performance Metrics



Questions?