# Temporal Anomaly Detection for Space Robotic Inspection

Lovely L. Andeo
*University of Toronto*
*Institute for Aerospace Studies*
Toronto, Canada
lovely.andeo@robotics.utias.utoronto.ca

Chang Won Lee
*University of Toronto*
*Institute for Aerospace Studies*
Toronto, Canada
john.lee@robotics.utias.utoronto.ca

Steven L. Waslander
*University of Toronto*
*Institute for Aerospace Studies*
Toronto, Canada
steven.waslander@robotics.utias.utoronto.ca

***Abstract*—Anomaly detection is critical for safe robotic proximity operations and autonomous spacecraft inspection. While single-frame detectors provide strong spatial localization, they do not leverage temporal coherence, and existing video anomaly detection methods do not generalize well to space imagery, where extreme illumination, large viewpoint variation, and limited data are persistent challenges. To address these challenges, we present STARS (Spatio-Temporal Anomaly Refinement System), a modular and training-efficient temporal refinement framework that augments a pretrained single-frame detector with learned temporal reasoning. Instead of training anomaly detection from scratch, STARS learns to refine detector predictions using short temporal context by fusing DINOv2-Large features with detector outputs through confidence-weighted residual correction. We also introduce the TALLO (Temporal Anomaly Localization in Lunar Orbit) dataset, the first annotated video benchmark for space anomaly detection, extending the ALLO single-frame benchmark with multi-frame sequences and pixel-accurate anomaly masks. On TALLO, STARS improves pixel-level AP by 13.3% and reduces FPR@95 by 13.6% over the single-frame FlowCLAS baseline, outperforming other existing single-frame and video anomaly detection methods. These results demonstrate that temporal refinement provides a practical and scalable path toward robust anomaly detection in space operations.**

***Index Terms*—anomaly detection, video analysis, space robotics, temporal modeling, foundation models, deep learning**

## I. Introduction

Canadarm2 is a robotic arm deployed on the International Space Station, supporting operations such as satellite servicing, spacecraft docking, and structural inspection. Its successor, Canadarm3, will operate on the Lunar Gateway where greater autonomy will be required due to communication delays between Earth and lunar orbit. In these settings, robotic systems must detect anomalous conditions such as structural damage, foreign object debris, and surface defects with minimal human intervention while maintaining a low false-positive rate, since spurious detections will disrupt operations [1]. Detecting these anomalies is particularly challenging under space-specific conditions, which include extreme illumination, rapidly changing viewpoints, and limited labeled anomaly data [2], [3].

Existing approaches fall into two categories. Single-frame detectors achieve strong spatial localization but do not exploit temporal coherence, making them sensitive to illumination changes across frames [4]–[6]. Video-based methods are designed to detect anomalies that manifest as irregular temporal events [7], [8] – actions or motions that deviate from normal dynamic – where as structural defects and foreign objects in the space imagery are static in nature, changing in appearance only due to shifting viewpoints and illumination rather than any motion of the anomaly itself.

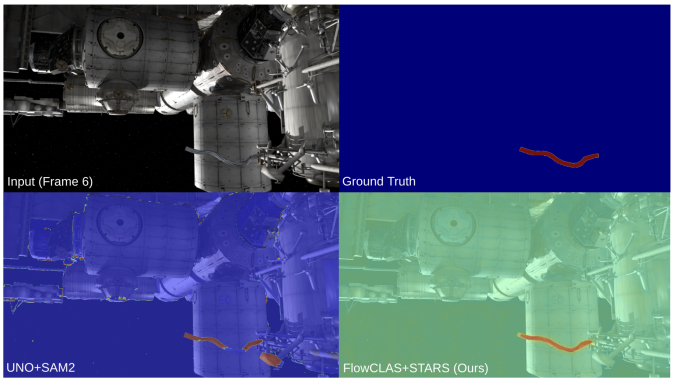


Fig. 1. Anomaly heatmaps on a challenging TALLO test frame where the anomaly blends into the background. FlowCLAS+STARS correctly localizes the anomaly while suppressing background activations, whereas the baseline approach, UNO+SAM2, produces incomplete segmentation and false negatives.

A practical alternative is to augment a strong single-frame detector with temporal context post-hoc. Heuristic approaches such as anomaly centroid tracking, Kalman filtering, and mask propagation via SAM2 [9] have been explored in this direction, but they assume that anomalies appear as discrete, spatially stable detections with reliable initial segmentations. In space imagery, this assumption often fails: specular reflections, shifting viewing angles, and illumination sweeps can cause anomaly signals to flicker, weaken, or become partially occluded across consecutive frames. When the initial prediction is uncertain, propagating it across the sequence can amplify errors rather than suppress them, as shown in Figure 1.

We address these problems with two contributions. First, we introduce **T**emporal **A**nomaly **L**ocalization in **L**unar **O**rbit (**TALLO**), a video benchmark that extends the ALLO dataset [2] with camera motion modeling under safety, cen-

tering, and collision avoidance constraints, creating multi-frame sequences and pixel-accurate ground-truth masks. Second, we present the **S**patio-**T**emporal **A**nomaly **R**efinement **S**ystem (**STARS**), a modular temporal adapter that augments a frozen single-frame detector with learned temporal reasoning. Instead of propagating heuristic segmentations, STARS fuses DINOv2 [10] features with frozen detector predictions and applies confidence-weighted residual refinement, focusing corrections on regions where the base detector is least certain. Our key contributions are:

- **TALLO**: the first annotated video benchmark for space anomaly detection simulating robotic inspection sweeps of the Canadarm. It contains 1320 sequences and 13,200 total frames, including 9 different anomalous objects.
- **STARS**: a lightweight temporal refinement adapter with 8.4M trainable parameters that augments a frozen single-frame detector without retraining the base model. STARS demonstrates state-of-the-art pixel-level performance on TALLO, outperforming single-frame, video-based, and heuristic temporal baselines across all primary metrics.

## II. Related Work

### A. Single-Frame Anomaly Detection

Recent progress in single-frame anomaly detection has largely converged on the use of frozen pretrained encoders paired with lightweight anomaly detection heads trained on normal samples. PatchCore [5] established a strong baseline on industrial anomaly detection benchmark MVTec AD [11] using coreset memory banks. Subsequent methods extended this paradigm through vision-language alignment [12], few-shot prompt adaptation [13], and unified DINOv2-based scoring [14]. DIAD [15] and InvAD [16] further explored hybrid reconstruction-discrimination objectives to improve multi-class anomaly detection. In the space domain, FlowCLAS [6] adapted this line of work to orbital imagery by combining DINOv2-Large features with normalizing flow contrastive learning, achieving state-of-the-art on the ALLO benchmark. In our experiments, we consider two additional representative general-domain baselines: UniNet [17], which uses student-teacher feature selection and contrastive learning to unify anomaly detection across diverse domains, and SuperSimpleNet [18], a discriminative model that generate synthetic anomalies in feature space using Perlin noise masks and combine segmentation head with a global classification head to reduce false positives. Collectively, these methods demonstrate that strong spatial anomaly localization can be learned from normal data alone, but they process each frame independently and therefore do not explicitly enforce temporal consistency under changing illumination and viewpoint. STARS builds on these single-frame foundations and adds a lightweight temporal adapter that learns to refine predictions across consecutive frames without retraining the base detector.

### B. Video Anomaly Detection

Video anomaly detection methods address the single-frame limitation by modeling temporal structure across frames. Prior work has explored reconstruction-based [19], [20], prediction based [21], [22], and representation-based [7], [8], [23], [24] frameworks, as well as pose-based normalizing flows [25] and multiple-instance learning [26]. These methods are developed primarily for surveillance scenarios where anomalies manifest as irregular actions or events in fixed-camera footage. Even when trained directly on a space imagery dataset like TALLO, video methods yield near-zero AP despite reasonable AUROC, as their coarse clip-level output lack the per-pixel spatial resolution that space inspection requires.

### C. Temporal Extensions for Single-Frame Methods

The most closely related prior works augment single-frame detectors with temporal post-processing rather than replacing them with fully video-native models. UNO [27] combines single-frame outlier detection with SAM2 [9] mask propagation, tracking anomaly centroids across frames to filter temporally inconsistent predictions; however, when initial segmentations are unreliable under harsh orbital lighting, this propagation amplifies rather than corrects errors.

Heuristic temporal extensions to single-frame detectors typically follow a detect-then-track paradigm. Simple centroid-based trackers link per-frame detections by proximity, while Kalman filter-based methods such as ByteTrack [28], SORT [**?**], and OC-SORT [29] add motion-model predictions and IoU-based association to handle brief occlusions. Applied to anomaly detection, these approaches gate or suppress detector scores outside tracked regions, effectively treating anomalies as persistent, spatially stable objects. This assumption is frequently violated in space imagery, where specular reflections and viewpoint changes cause anomaly evidence to flicker or degrade across consecutive frames. STARS addresses this by replacing heuristic propagation rules with learned temporal corrections gated by per-pixel detector confidence.

## III. Methodology

STARS augments a frozen single-frame detector with short-range temporal reasoning through four stages: (1) single-frame detection, (2) feature-anomaly fusion, (3) temporal modeling, and (4) confidence-gated residual refinement (Figure 2). Only stages 2–4 are trained. The base detector and DINOv2 encoder remain frozen throughout.

### A. Single-Frame Anomaly Detection With FlowCLAS

We adopt the pretrained FlowCLAS [6] as our base single-frame anomaly detector, as it represents the current state-of-the-art for anomaly detection in space imagery. This frozen detector processes each frame independently to produce a normalized anomaly map

$$\hat{\mathcal{M}}_{\text{det}} \in [0,1]^{H \times W} \tag{1}$$

and a frame-level score $s_{\text{det}}$. Keeping the detector frozen ensures that STARS learns to complement rather than override its predictions, and that the adapter is a plug-and-play module that can potentially be used with different base detectors – a direction that we will explore in the future.

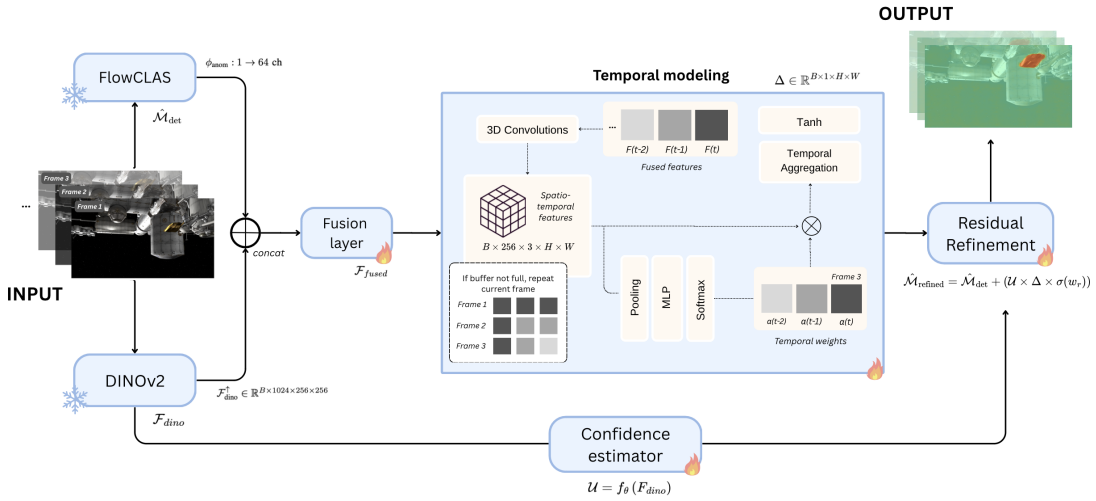


Fig. 2. Overview of the STARS pipeline. Consecutive input frames are processed by a frozen FlowCLAS detector and a frozen DINOv2 encoder, producing anomaly maps $\hat{\mathcal{M}}_{\text{det}} \in [0,1]^{H\times W}$ and patch features $\mathcal{F}_{\text{dino}}$ respectively. The anomaly map is projected via $\phi_{\text{anom}}$ and concatenated with upsampled $\mathcal{F}_{\text{dino}}$ before a trainable fusion layer produces $\mathcal{F}_{\text{fused}} \in \mathbb{R}^{B\times256\times256\times256}$. A temporal module processes a window of $w$=3 consecutive fused maps via 3D convolutions and content-driven attention – which assigns lower weight to frames degraded by lighting or occlusion – to produce per-pixel residual corrections $\Delta \in [-1,1]^{B\times1\times H\times W}$. A confidence estimator $\mathcal{U}$ gates corrections toward uncertain regions, yielding $\hat{\mathcal{M}}_{\text{refined}}$, where $\sigma(w_r) \leq 0.5$ limits the correction magnitude. Trainable components are marked with fire icon.

### B. Feature-Anomaly Map Fusion

Raw detector scores identify where anomalies are likely but carry no semantic context. We extract patch-level features from a frozen DINOv2-Large encoder, whose structural and textural representations remain stable under extreme illumination variation, separating genuine anomaly responses from lighting-induced artifacts. We upsample $\mathcal{F}_{dino}$ to $256 \times 256$, balancing spatial detail against downstream memory cost. The anomaly map $\hat{\mathcal{M}}_{det}$ s projected from 1 to 64 channels via a Conv-BN-ReLU head $\phi_{anom}$ preventing it from being overwhelmed by the 1024-dimensional DINOv2 features upon concatenation. The fusion layer then reduces the 1088-channel concatenation to 512 then 256 channels, producing a joint encoding of semantic content and per-pixel anomaly evidence.

### C. Temporal Modeling

Temporal context is necessary to accumulate consistent evidence across frames and reduce sensitivity to per-frame lighting conditions. The fused feature maps $\mathcal{F}_{\text{fused}}$, where the 256-channel dimension is the output of the fusion layer, serve as input to the temporal module. Stacking $w = 3$ consecutive maps gives:

$$\mathcal{T} \in \mathbb{R}^{B\times256\times w\times256\times256}. \tag{2}$$

These are processed through two 3D convolutional blocks that jointly encode spatial structure and cross-frame dependencies, progressively reducing the channel dimension via a bottleneck design. A content-driven temporal attention mechanism assigns per-frame scalar weights $\alpha_t \in \mathbb{R}^w$, derived via adaptive pooling and learned linear transformations. Rather than applying a fixed recency bias, this mechanism selectively down-weights frames in which the anomaly signal is degraded due to lighting or occlusion. The attended features are aggregated by a $w\times1\times1$ convolution and decoded by a spatial refiner with tanh activation into per-pixel residual corrections:

$$\Delta \in [-1,1]^{B\times1\times H\times W}, \tag{3}$$

The window size $w = 3$ is kept deliberately compact: it is sufficient for resolving frame-to-frame lighting flicker and accumulating short-range evidence, while remaining computationally tractable for onboard hardware.

### D. Confidence-Weighted Refinement

Applying residual corrections uniformly risks degrading regions where the base detector is already accurate. We instead estimate a per-pixel confidence map

$$\mathcal{U} = f_\theta(\mathcal{F}_{\text{dino}}) \in [0,1]^{H\times W}, \tag{4}$$

where $\mathcal{F}_{dino} \in \mathbb{R}^{B\times1024\times H_p\times W_p}$ are the DINOv2 patch features upsampled to match the target resolution before the lightweight convolutional head $f_\theta$, $\mathcal{U}$ is the highest regions where $\hat{\mathcal{M}}_{det}$ is uncertain and lowest where it is reliable, so corrections are concentrated where they are most needed. Residual corrections are gated spatially:

$$\hat{\mathcal{M}}_{\text{refined}} = \hat{\mathcal{M}}_{\text{det}} + \sigma(w_r) \cdot \Delta \cdot \mathcal{U}, \tag{5}$$

where the scalar $w_r$ is learned and bounded by 0.5 through a sigmoid reparameterization, which biases the model toward conservative residual corrections and limits deviations from the base detector prediction.

### E. Training Objectives

STARS is trained with three complementary losses operating at the sequence, frame, and pixel levels respectively.

*Sequence-Level Contrastive Loss.* A dual triplet loss on spatially-pooled sequence representations separates normal and anomalous sequences in feature space:

$$\mathcal{L}_{\text{contrast}} = \sum_{i=1}^{2} \log(1 + \exp(d(a, p_i) - d(a, n) + m)), \quad (6)$$

where $a$, $p_i$, $n$ denote the anchor (normal), hard positive, and anomalous embeddings, and $m$=1.0. Operating on pooled rather than dense spatial representations makes this loss memory-efficient and robust to the spatial variability of anomaly appearance across frames.

*Frame-Level Temporal Consistency.* Inter-frame displacement in TALLO is bounded at $\leq 0.4$m per frame, small enough that the underlying scene structure remains largely consistent between consecutive frames. Under this constraint, abrupt changes in $\mathcal{F}_{fused}$ are more likely to reflect illumination artifacts than genuine structural changes, making smoothness a reasonable prior. We penalize the cosine distance between temporally adjacent fused representations, the output of the fusion layer at frame $t$:

$$\mathcal{L}_{\text{temp}} = \mathbb{E}_t \left[ 1 - \frac{\mathcal{F}_t \cdot \mathcal{F}_{t-1}}{\|\mathcal{F}_t\|_2 \|\mathcal{F}_{t-1}\|_2} \right], \quad (7)$$

which encourages smooth feature evolution as the camera sweeps around the spacecraft.

*Pixel-Level Correction Loss.* Rather than supervising the refined output with binary cross-entropy alone, we add correction-aware terms that reward error reduction relative to the baseline and penalize regressions in correctly predicted regions:

$$\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{BCE}}(y, \hat{\mathcal{M}}_{\text{refined}}) + \alpha \mathcal{L}_{\text{improve}} + \beta \mathcal{L}_{\text{degrade}}, \quad (8)$$

where $\mathcal{L}_{\text{improve}}$ penalizes pixels where the refined score moves away from the ground truth relative to $\hat{\mathcal{M}}_{\text{det}}$, and $\mathcal{L}_{\text{degrade}}$ penalizes pixels where the refined score regresses on an already correct base prediction. This encourages the residual head to fill in gaps left by the base detector rather than rewrite its predictions entirely, which is particularly important when the base detector is already strong. The total loss is:

$$\mathcal{L} = \lambda_c \mathcal{L}_{\text{contrast}} + \lambda_t \mathcal{L}_{\text{temp}} + \lambda_{\text{corr}} \mathcal{L}_{\text{corr}}, \quad (9)$$

with $\lambda_c$=0.5, $\lambda_t$=0.3, $\lambda_{\text{corr}}$=1.0, $\alpha$=0.5, $\beta$=1.0.

## IV. TALLO Dataset

TALLO extends ALLO [2], a photorealistic single-frame dataset of the ISS rendered in Blender with physically accurate lighting and camera poses modelling Canadarm's operational positions. While ALLO provides spatially diverse static viewpoints, it lacks the temporal continuity required to train or evaluate video-based anomaly detection methods. We address this dataset gap by introducing camera motion into the same Blender scene to generate multi-frame sequences with pixel-accurate ground-truth masks while preserving ALLO's rendering methodology.

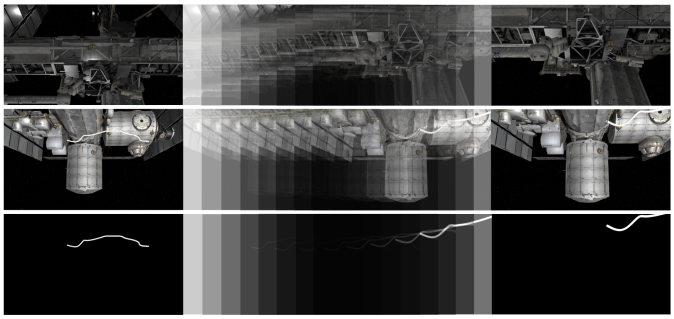

Fig. 3. Representative sequences from the TALLO dataset. *Top*: a clean ISS sequence. *Middle*: an anomalous sequence. *Bottom*: corresponding ground-truth anomaly masks.

*Camera Design and Inspection Target*. Each sequence is rendered from one of 41 predefined camera positions distributed across the ISS structure, with disjoint subsets assigned to training (cameras 14–35), validation (cameras 36–39), and test (cameras 0–13). Camera positions are derived from ALLO's original viewpoint set, with select positions adjusted to preserve adequate ISS coverage when camera motion is introduced. This split ensures that no viewpoint seen during training or validation appears at test time, preventing the model from exploiting memorized perspective cues. Each camera position is paired with a corresponding grapple fixture location, which serves as the designated inspection target. All camera trajectories are constrained to maintain a minimum standoff distance of $>$1m from the grapple fixture, with collision avoidance applied at each frame to prevent intersection with the ISS structure or other scene objects.

*Sequence Generation*. TALLO produces two categories of 10-frame sequences rendered at 1920×1080: *clean sequences* of the nominal ISS structure, and *anomalous sequences* containing authentic ALLO objects including thermal blankets, cables, satellites, and handheld tools. Anomalies are placed within the camera frustum at a fixed world position for the duration of each sequence, with orientation randomized to avoid edge-on projections. Pixel-accurate ground-truth masks are generated via Blender's compositor using pass indices for ISS components (2–8) and anomalies (9).

Training and validation sequences follow *random smooth trajectories* in which each frame's displacement is drawn stochastically but biased toward the paired grapple fixture, with the bias weight increasing linearly over the sequence so that later frames tend to orbit the inspection target. Step directions are perturbed with zero-mean Gaussian noise and clamped to a maximum per-frame displacement of $\leq$0.4m, producing smooth but unpredictable paths that expose the model to diverse viewpoint transitions without repeating any fixed pattern. Test sequences follow *deterministic arc trajectories* in which the camera sweeps along a circular arc of fixed radius (5m) and angular span (60°) centered on the grapple fixture, in one of four cardinal directions (left, right, up, down), yielding $14$ cameras $\times$ $4$ directions $=$ $56$ distinct test trajectories per sequence type.

This motion regime mismatch is intentional: stochastic training trajectories prevent memorization of fixed motion patterns, while deterministic arc sweeps at test time evaluate generalization to the structured camera paths characteristic of teleoperated inspection – consistent with Canadarm operational practice. Inter-frame displacement is bounded at $\leq 4$m, keeping the ISS within the camera frustum across frames. At 10 frames per sequence, this yields a maximum swept distance of $\approx 3.6$m, sufficient to capture the parallax and illumination variation that temporal methods must handle.

*Statistics*. TALLO contains 880 training sequences (22 camera positions), 160 validation sequences (4 positions), and 280 test sequences (14 positions, arc trajectories), split evenly between clean and anomalous. ShapeNet objects are injected into half the training and validation anomalous sequences as hard negatives for contrastive learning, with model IDs partitioned to be disjoint across all splits and from test-set ALLO anomalies, enforcing a strict generalization challenge. Examples of the rendered sequences are shown in Figure 3.

## V. Experiments

### A. Implementation Details

We implement STARS in PyTorch 2.0 and train for up to 50 epochs using AdamW with learning rate $5\times10^{-5}$, weight decay 0.01, and effective batch size 4 (batch 2 with 2-step gradient accumulation). Early stopping is applied with patience 10 using validation AUPRC as the stopping criterion. Mixed-precision training (FP16) is used throughout. To stabilize early optimization, the temporal consistency loss weight $\lambda_t$ is linearly ramped during the first 2 epochs. All 10 frames in each sequence are supervised under a fixed temporal window of $w$=3. During training, sequences are presented in a fixed order to preserve the alternating normal/anomalous pairing required by the contrastive loss; each sequence contributes all 10 frames as a single batch element. FlowCLAS processes each frame at its native resolution; DINOv2 resizes frames to $224\times224$ for feature extraction, and all fused representations are computed at $256\times256$.

Only the fusion layers, temporal module, confidence estimator, and refinement head are updated, for a total of 8.4M trainable parameters compared to the frozen base detector with 127M parameters. Training requires approximately 48 hours on a single NVIDIA A100 (40GB). The model trains exclusively on nominal ISS configurations; ShapeNet [30] outlier injections provide hard negatives for contrastive learning, and test sequences contain only authentic ALLO anomalies never seen during training.

### B. Evaluation Metrics

We evaluate pixel-level anomaly localization using four standard metrics. *AUROC* measures threshold-independent discrimination. *AP* summarizes precision-recall performance and is sensitive to the quality of high-confidence predictions. *AUPRO* evaluates per-region localization quality via connected-component overlap at FPR $\leq 0.3$, rewarding methods that recover the full spatial extent of an anomaly. *FPR@95* measures false-positive burden at high recall, which is critical in space robotics where false alarms trigger costly operational halts. We implement 5000-bin histogram-based metric accumulation to reduce evaluation memory overhead relative to full-tensor accumulation.

### C. Comparison with Other Methods

TABLE I
Comparison of methods on TALLO test set.

| Method | AUROC↑ | AP↑ | FPR@95↓ | AUPRO↑ |
|---|---|---|---|---|
| *Single-frame and video methods* | | | | |
| UniNet [17] | 0.8946 | 0.1230 | 0.8946 | 0.5965 |
| SuperSimpleNet [18] | 0.7323 | 0.0706 | 0.8798 | 0.3457 |
| PEL4VAD [8] | 0.6600 | 0.0168 | 0.7519 | 0.1543 |
| BN-WVAD [7] | 0.6664 | 0.0160 | 0.6942 | 0.2197 |
| *Single-frame detectors with temporal extensions* | | | | |
| UNO [27] | 0.7081 | 0.0862 | 0.8706 | 0.6210 |
| UNO+SAM2 | 0.6606 | 0.0629 | 0.8815 | 0.6507 |
| UNO† | 0.8536 | 0.4707 | 0.8271 | 0.6948 |
| UNO+SAM2† | 0.8296 | 0.1564 | 0.8311 | 0.7145 |
| FlowCLAS | 0.9493 | 0.6970 | 0.3517 | **0.7523** |
| *FlowCLAS + STARS (Ours)* | **0.9601** | **0.7895** | **0.3040** | 0.7122 |

*1) Single-Frame and Video Methods:* Table I compares STARS against representative single-frame and video baselines under the same training and evaluation protocol. Video methods (PEL4VAD, BN-WVAD) transfer poorly to TALLO, suggesting that temporal modeling alone is insufficient without inductive biases suited to space inspection imagery. Among the single-frame baselines, UniNet's large AUROC–AP gap indicates weak precision at high-confidence thresholds, likely due to domain shift.

*2) Single-Frame Detectors with Temporal Extensions:* Table I also reveals two structural findings from comparisons with an existing single-frame detector with temporal extension. First, base detector quality dominates: UNO performs poorly pretrained on Cityscapes and improves substantially after fine-tuning (denoted by †), confirming that domain adaptation is a prerequisite for meaningful performance on TALLO. Second, SAM2-based propagation consistently degrades performance in both settings, suggesting that under extreme orbital lighting, errors in the initial segmentation are more often propagated than corrected across frames, as observed in Figure 1. FlowCLAS provides a considerably stronger foundation; applying STARS yields consistent gains across AUROC, AP, and FPR@95, and achieves the strongest AP across all methods by a wide margin, demonstrating that augmenting a strong single-frame detector with learned temporal refinement is more effective on TALLO than either spatial-only baselines, surveillance-domain video methods, or with SAM2 tracking extensions. The decrease in AUPRO suggests a more conservative refinement behavior that suppresses false positives at the cost of slightly tighter anomaly regions.

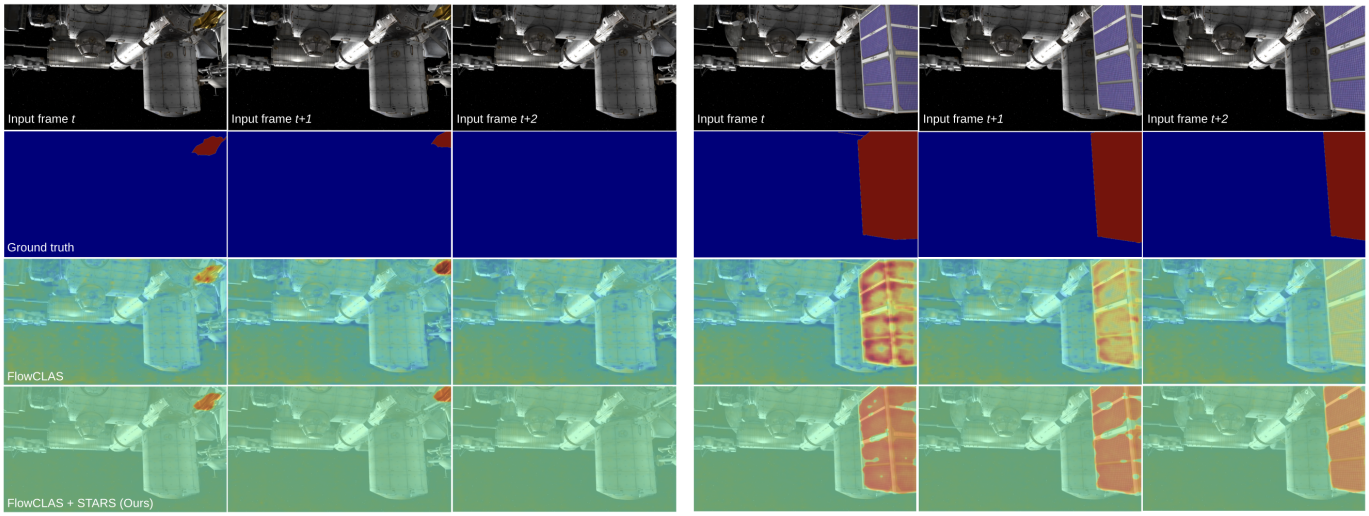


Fig. 4. Qualitative comparison on two TALLO test sequences (three representative frames each). *Left*: a thermal blanket anomaly. FlowCLAS produces a diffuse heatmap that degrades as the camera angle changes; STARS sharpens the anomaly boundary across frames and suppresses background activations once the blanket exits view. *Right*: a satellite occluding the ISS. FlowCLAS detects object edges but misses the interior and loses localization on the third frame; STARS recovers full object extent across all frames by down-weighting the degraded frame via temporal attention.

### D. Ablation Studies

Table II progressively adds each STARS component to the base detector. *Temporal modeling* alone slightly degrades performance: without semantic context, the temporal module may treat lighting-induced score changes the same way it treats genuine anomaly signals, and the resulting corrections may result to worsened predictions. *DINOv2 features* recover and surpass the baseline across all primary metrics, providing illumination-invariant structural context for cross-frame comparison and grounding the confidence estimator. *Confidence weighting* delivers further gains by gating corrections to uncertain regions, preventing the residual head from overwriting reliable predictions.

TABLE II
ABLATION STUDY ON TALLO TEST SET.

| Image Detector | Temp Modeling | DINOv2 | Conf. Weighting | AUROC↑ | AP↑ | FPR@95↓ | AUPRO↑ |
|---|---|---|---|---|---|---|---|
| ✓ | | | | 0.9493 | 0.6970 | 0.3517 | **0.7523** |
| ✓ | ✓ | | | 0.9432 | 0.6635 | 0.4353 | 0.6972 |
| ✓ | ✓ | ✓ | | 0.9534 | 0.7228 | 0.3352 | 0.7508 |
| ✓ | ✓ | ✓ | ✓ | **0.9601** | **0.7895** | **0.3040** | 0.7122 |

Table III holds the base detector fixed at FlowCLAS to isolate the effect of each temporal extension. Tracking-based methods (centroid and Kalman) assume spatially stable anomalies, an assumption frequently violated in TALLO, and consequently degrade AP substantially. SAM2 performs worst on FPR@95, reinforcing poor initial segmentations across frames under harsh lighting rather than correcting them. STARS achieves the largest AP gain among all extensions, demonstrating that learned refinement reshapes the anomaly score distribution rather than merely smoothing it. Qualitative results are shown in Figure 4.

TABLE III
FLOWCLAS WITH DIFFERENT TEMPORAL EXTENSIONS.

| Method | AUROC↑ | AP↑ | FPR@95↓ | AUPRO↑ |
|---|---|---|---|---|
| FlowCLAS | 0.9493 | 0.6970 | 0.3517 | **0.7523** |
| + Centroid tracking | 0.9283 | 0.3013 | 0.3812 | 0.6696 |
| + Kalman filter | 0.9333 | 0.5043 | 0.3010 | 0.6982 |
| + SAM2 | 0.8277 | 0.5512 | 0.8392 | 0.6480 |
| +*STARS (Ours)* | **0.9601** | **0.7895** | **0.3040** | 0.7122 |

## VI. CONCLUSION

We presented STARS, a modular spatio-temporal adapter that equips a frozen single-frame anomaly detector with learned temporal reasoning for space inspection. Across all major comparisons on TALLO, STARS improves pixel-level anomaly localization over strong single-frame baselines, surveillance-domain video anomaly detection methods, and heuristic temporal extensions. STARS achieves these gains with only 8.4M trainable parameters, making temporal refinement a practical alternative to retraining fully video-native models. We also introduced TALLO, the first temporal benchmark for space anomaly detection, enabling controlled evaluation under viewpoint change and severe illumination variation. An important next step is evaluating the framework on real ISS imagery to quantify the simulation-to-real gap. Future work will also examine longer temporal contexts and more complex operational settings. More broadly, our results indicate that lightweight temporal refinement may be an effective strategy for anomaly detection in other challenging inspection domains.

## ACKNOWLEDGMENT

The authors thank MDA Space for their support of the preceding work that laid the foundation for this research, and for their continued engagement on this work through feedback and discussion.

## REFERENCES

[1] M. Smith *et al.*, "An overview of NASA's activities to return humans to the moon and establish a sustainable presence in orbit and on the surface," in *Proc. IEEE Aerospace Conf.*, Mar. 2020, pp. 1–10.

[2] S. Leveugle, C. Lee, S. Stolpner, C. Langley, P. Grouchy, S. Waslander, and J. Kelly, "A photorealistic dataset and vision-based algorithm for anomaly detection during proximity operations in lunar orbit," arXiv:2409.20435, Sep. 2024.

[3] T. Uriot, D. Izzo, L. F. Simões, R. Abay, N. Einecke, S. Rebhan, J. Martinez-Heras, F. Letizia, J. Siminski, and K. Merz, "Spacecraft collision avoidance challenge: Design and results of a machine learning competition," *Astrodynamics*, vol. 8, pp. 103–120, 2024.

[4] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, 2022, pp. 98–107.

[5] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 318–14 328.

[6] C. W. Lee, S. Leveugle, P. Grouchy, C. Langley, S. Stolpner, J. Kelly, and S. L. Waslander, "FlowCLAS: Enhancing normalizing flow-based anomaly segmentation via contrastive learning," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, 2026, pp. 6998–7007.

[7] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. Shen, "BatchNorm-based weakly supervised video anomaly detection," arXiv:2311.15367, 2023.

[8] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning prompt-enhanced context features for weakly supervised video anomaly detection," *IEEE Trans. Image Process.*, vol. 33, pp. 4923–4936, Sep. 2024.

[9] N. Ravi *et al.*, "SAM 2: Segment anything in images and videos," arXiv:2408.00714, Aug. 2024.

[10] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Trans. Machine Learning Research (TMLR)*, 2024.

[11] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 9524–9534.

[12] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 606–19 616.

[13] Z. Li *et al.*, "PromptAD: Learning prompts with only normal samples for few-shot anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] J. H. Lee and C. Kim, "UniFormaly: Towards task-agnostic unified framework for visual anomaly detection," arXiv:2307.12325, 2023.

[15] Z. He, Z. Liu, Y. Li, and K. Wang, "DiAD: A diffusion-based framework for multi-class anomaly detection," in *Proc. AAAI Conf. Artificial Intelligence*, 2024.

[16] Y. Liu, C. Zhang, and L. Zheng, "Revisiting reverse distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[17] S. Wei, J. Jiang, and X. Xu, "UniNet: A contrastive learning-guided unified framework with feature selection for anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2025.

[18] B. Rolih, M. Fučka, and D. Skočaj, "No label left behind: A unified surface defect detection model for all supervision regimes," *J. Intelligent Manufacturing*, 2025.

[19] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 733–742.

[20] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1705–1714.

[21] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection: A new baseline," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6536–6545.

[22] M. Ye, X. Peng, W. Gan, W. Wu, and Y. Qiao, "AnoPCN: Video anomaly detection via deep predictive coding network," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2019, pp. 1805–1813.

[23] Z. Chen, W. Liu, and X. Xie, "TEVAD: Improved video anomaly detection with captions," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

[24] A. Wu, S. Tang, B. Jiang, X. Guo, J. Wang, and T. Lu, "VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artificial Intelligence*, 2024.

[25] O. Hirschorn and S. Avidan, "Normalizing flows for human pose anomaly detection," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 13 545–13 554.

[26] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8022–8031.

[27] A. Delić, M. Grčić, and S. Šegvić, "Outlier detection by ensembling uncertainty with negative objectness," in *Proc. British Machine Vision Conf. (BMVC)*, 2024, paper no. 779.

[28] Y. Zhang *et al.*, "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. European Conf. Computer Vision (ECCV)*, 2022, pp. 1–21.

[29] J. Cao, X. Peng, and J. Peng, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9686–9696.

[30] A. X. Chang *et al.*, "ShapeNet: An information-rich 3D model repository," arXiv:1512.03012, Dec. 2015.