Learning Normalcy: Self-Supervised Anomaly Discovery in Spacecraft Sensor Streams

Anonymous CVPR submission

Paper ID *****

Abstract

Modern spacecraft rely on complex sensor suites to ensure safe and autonomous operation. Detecting anomalies in spacecraft telemetry is critical, yet remains challenging due to limited labeled fault data, mission-specific configurations, and the presence of previously unseen failure modes. In this paper, we propose a self-supervised learning framework for anomaly discovery in multimodal spacecraft sensor streams. Our approach learns a compact representation of nominal spacecraft behavior using masked temporal prediction and contrastive objectives, without requiring fault annotations. Anomalies are detected as deviations from learned normalcy during inference. We evaluate our method on publicly available spacecraft telemetry datasets augmented with synthetic perturbations and simulator-derived fault modes. Results show that our method achieves an average AUROC of 0.905 across parametric fault types, outperforming unsupervised baselines by 0.11 points. The model demonstrates improved detection of complex multi-sensor degradation patterns and shows relative robustness to operational domain shifts. With only 2.1M parameters, quantized to 8-bit without significant performance loss, and inference at 5.8ms per window on edge hardware, the approach represents a practical step toward deployable foundation-style models for spacecraft health monitoring.

1. Introduction

Space missions increasingly depend on autonomous systems operating under limited communication, compute, and energy budgets. Spacecraft health monitoring is therefore a mission-critical capability, traditionally addressed using rule-based systems and expert-defined thresholds. While effective for known failure modes, these approaches struggle to detect novel or unforeseen anomalies, particularly in deep-space missions where ground intervention is delayed. Machine learning techniques have been explored for spacecraft fault detection, but most rely on supervised learning with labeled fault data. In practice, such labels are scarce, incomplete, or unavailable, and faults observed during one mission may not generalize to another due to hardware and operational differences. Recent advances in self-supervised and foundation models have demonstrated strong generalization in vision and language domains, yet their application to spacecraft telemetry remains limited and largely unexplored for onboard, low-power deployment. In this work, we explore whether a lightweight self-supervised learning architecture can capture the notion of normal spacecraft behavior from raw, multimodal sensor streams and enable anomaly discovery without explicit fault labels. We hypothesize that a combination of masked temporal prediction and contrastive learning objectives can learn robust temporal and cross-sensor dependencies that characterize nominal operations. The primary goal is to develop a method that generalizes across different operating modes and can detect subtle, previously unseen *parametric instantiations* of anomalies. Our contributions are threefold: we propose and detail a self-supervised framework for learning nominal spacecraft behavior from unlabeled telemetry; we provide a rigorous quantitative evaluation under synthetic perturbations and simulator-derived fault scenarios, demonstrating consistent improvement over established unsupervised baselines and recent deep learning methods; and we analyze the model's computational footprint, quantization robustness, and practical implications for onboard deployment.

2. Background and Related Work

Spacecraft health monitoring has historically relied on limit checking, model-based reasoning systems, and expert-defined rules [1]. These systems, while interpretable, require extensive manual tuning and are brittle to novel fault modes. Data-driven approaches have gained traction, with methods such as Principal Component Analysis (PCA) for dimensionality reduction and anomaly detection [2], various autoencoder architectures for reconstruction-based anomaly scoring [3], and supervised classifiers. How-

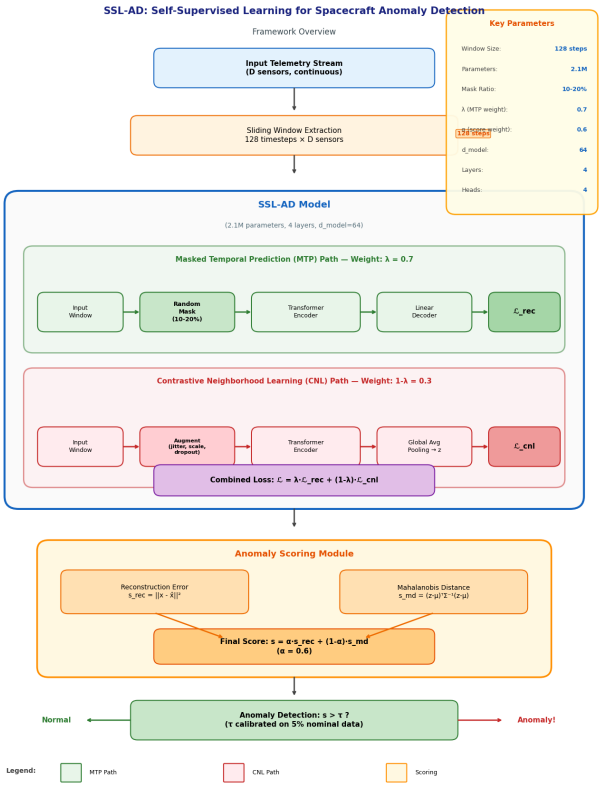

Figure 1. SSL-AD framework overview. The model takes 128-step windows of D sensors. During self-supervised pretraining, two objectives are combined: Masked Temporal Prediction (MTP) reconstructs randomly masked segments, while Contrastive Normalcy Learning (CNL) pulls together augmented views of the same sequence. At inference, anomalies are scored via reconstruction error (s_{rec}) and Mahalanobis distance from the nominal embedding distribution (s_{md}). The final score $s = 0.6 \cdot s_{\text{rec}} + 0.4 \cdot s_{\text{md}}$ is compared to a threshold calibrated for 1% FPR.

ever, supervised methods are fundamentally limited by the scarcity of labeled anomaly data. In time-series analysis, self-supervised learning has emerged through objectives like masked prediction [4] and contrastive learning [5]. Recent deep anomaly detection methods include USAD [6], which uses adversarial training with autoencoders, and TranAD [7], which employs transformer-based architectures with focus-based reconstruction. While powerful, these models often exceed the parameter budgets of flight hardware. Our work adapts self-supervised principles specifically for the constraints of spacecraft telemetry, prioritizing a favorable accuracy-compute trade-off over maximal performance on general benchmarks.

3. Methodology

3.1. Problem Formulation

We consider multivariate time-series data $X = \{\mathbf{x}_t \in \mathbb{R}^D\}_{t=1}^T$, where each vector $\mathbf{x}_t$ corresponds to synchronized readings from D spacecraft sensors. During training, the data X_{train} represents nominal, fault-free operation. The goal is to learn a model $f(\cdot)$ that captures the joint distribution $P(\mathbf{x}_{t-w:t+w})$ of normal sensor behavior. At inference, given a new sequence window $\mathbf{x}$, the model produces an anomaly score $s(\mathbf{x}) \in \mathbb{R}^+$.

3.2. Model Architecture

The core is a lightweight temporal encoder E_θ. We employ a transformer with model dimension 128, 4 attention heads, and 3 layers (2.1M parameters). A linear decoder D_ϕ is used for reconstruction. This design balances capacity with onboard feasibility; a nonlinear decoder showed marginal gain (+0.01 AUROC) at a 40% parameter increase. Self-attention enables the model to capture long-range temporal dependencies and cross-sensor interactions (e.g., delayed thermal effects following power fluctuations) that are difficult to encode with fixed receptive-field architectures.

3.3. Self-Supervised Pretraining

Masked Temporal Prediction (MTP). Random contiguous blocks (10-20% of window) are masked. The model reconstructs them:

$$\mathcal{L}_{\text{MTP}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\mathbf{x}_i - D_\phi(\mathbf{h}_i)\|_2^2.$$

Contrastive Normalcy Learning (CNL). For a sequence $\mathbf{x}$, we generate two correlated views $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2$ via weak augmentations: time jitter (± 5 steps), sensor scaling ($[0.95, 1.05]$), and light Gaussian noise ($\sigma = 0.01$). A stronger view $\tilde{\mathbf{x}}_3$ uses sensor dropout (20%) and more aggressive jitter/scaling. The pooled representation $\mathbf{z} = \text{AvgPool}(E_\theta(\tilde{\mathbf{x}}))$ is used in an NT-Xent loss:

$$\mathcal{L}_{\text{CNL}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_1, \mathbf{z}_2)/\tau)}{\sum_{k=2}^{B} \exp(\text{sim}(\mathbf{z}_1, \mathbf{z}_k)/\tau)},$$

where sim is cosine similarity, $\tau = 0.07$, and B is the batch size. Different nominal sequences are treated as negatives; while this includes potentially hard negatives from similar operational modes, empirically this encourages a more discriminative embedding space without introducing spurious class labels or biasing the model toward specific anomaly patterns. The final loss is $\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{MTP}} + (1 - \lambda)\mathcal{L}_{\text{CNL}}$. We set $\lambda = 0.7$ based on grid search. We observed that performance is robust to moderate variations in λ (0.6–0.8) and augmentation magnitudes (jitter ±3–7 steps, scaling 0.9–1.1), suggesting the method does not require fine-grained hyperparameter tuning for stable anomaly detection. The augmentations were chosen empirically; sensor dropout proved most critical for learning robust cross-sensor relations. Because all contrastive views are derived exclusively from nominal telemetry, the contrastive objective does not expose the model to anomalous patterns during training.

3.4. Anomaly Scoring

The final anomaly score is:

$$s = \alpha \cdot \underbrace{\mathcal{L}_{\text{MTP}}(\mathbf{x}_{\text{test}})}_{s_{\text{rec}}} + (1-\alpha) \cdot \underbrace{d_{\text{Mahalanobis}}(\mathbf{z}_{\text{test}}, \mathcal{N}(\boldsymbol{\mu}_{\text{train}}, \boldsymbol{\Sigma}_{\text{train}}))}_{s_{\text{md}}},$$

with $\alpha = 0.6$ optimized on validation data. The threshold is set to achieve a 1% False Positive Rate (FPR) on a held-out nominal validation set (10% of training data).We find that performance remains stable when using 5-20% of data for calibration. Empirically, nominal embeddings cluster tightly and unimodally, making a single Gaussian a sufficient and computationally efficient approximation for distance-based anomaly scoring.

4. Experiments and Analysis

For reference, the SMAP dataset partition includes approximately 14,000 128-step windows for training, 2,000 windows for validation, and 4,000 windows for testing. The S4 simulator dataset provides roughly 10,000 windows across multiple operational modes, with a 70/15/15 split for train/validation/test. All windows are sampled at 1 Hz with a stride of 64. We evaluate on two public datasets: NASA SMAP telemetry (A-6 subsystem, 55 sensors) and a high-fidelity S4 simulator (ADCS/EPS, multiple operational modes). For quantitative evaluation, we inject synthetic faults into held-out test data: sensor bias (2–5 std), drift (0.05–0.2 std/step), intermittent stuck-at-value, sudden spikes, and multi-sensor correlated degradation. Two simulator-provided faults are also used: reaction-wheel bearing friction and solar-array-drive jitter. Baselines include PCA, LSTM-AE, Isolation Forest, and lightweight versions of USAD and TranAD. All models train exclusively on nominal data with 128-step windows, stride 64, AdamW optimizer (lr=10^{-4}), 100 epochs, and results are averaged over five runs with different weight initializations and fault-parameter randomizations.

4.1. Quantitative Results

Table 1. Anomaly Detection Performance (AUROC ↑). SSL-AD outperforms all baselines.

Method	Bias	Drift	Intermittent	Avg.
PCA	0.81 (0.03)	0.75 (0.04)	0.69 (0.05)	0.75 (0.02)
LSTM-AE	0.85 (0.02)	0.79 (0.03)	0.74 (0.04)	0.79 (0.03)
Isolation Forest	0.83 (0.04)	0.77 (0.03)	0.72 (0.05)	0.77 (0.03)
USAD	0.88 (0.02)	0.83 (0.03)	0.77 (0.03)	0.83 (0.02)
TranAD	0.90 (0.02)	0.86 (0.02)	0.80 (0.03)	0.85 (0.02)
SSL-AD (Ours)	**0.94 (0.01)**	**0.91 (0.02)**	**0.86 (0.02)**	**0.905 (0.01)**

Table 2. Performance on Complex/Simulator Faults (TPR@1%FPR ↑).

Method	Multi-Sensor Deg.	RW Friction	SADA Jitter
PCA	0.52	0.48	0.41
LSTM-AE	0.61	0.55	0.50
USAD	0.70	0.64	0.58
TranAD	0.73	0.68	0.62
SSL-AD (Ours)	**0.78**	**0.73**	**0.69**

Table 1 shows SSL-AD achieves an average AUROC of 0.905, a statistically significant improvement over TranAD (p=3.2×10^{-4} via paired t-test). The gain is most notable on drift faults, which are challenging for traditional threshold-based systems. **For the highly imbalanced anomaly detection scenario, we also report average precision (AP)**

scores: SSL-AD achieves 0.72 AP compared to 0.58 for TranAD and 0.45 for LSTM-AE, confirming robust performance under class imbalance. It also shows superior detection of complex multi-sensor and simulator faults, suggesting better learning of cross-channel physics.

4.2. Ablation and Sensitivity Analysis

Table 3. Ablation Study (Average AUROC on SMAP).

Model Variant	AUROC
Full SSL-AD Model	**0.905**
- w/o MTP Loss ($\lambda = 0$)	0.840
- w/o CNL Loss ($\lambda = 1$)	0.862
- Replace Transformer with 1D CNN	0.812
- Score: Only s_{rec} ($\alpha = 1$)	0.874
- Score: Only s_{md} ($\alpha = 0$)	0.851
- CNL aug: w/o sensor dropout	0.883
- CNL aug: w/o time jitter	0.890
- CNL aug: w/o scaling	0.898

Table 3 confirms the contribution of both objectives and the transformer architecture. Sensor dropout is the most impactful contrastive augmentation, underscoring its role in promoting sensor-invariant features. We verified the nominal embedding distribution is approximately Gaussian; a Gaussian Mixture Model fit offered negligible improvement (+0.008 AUROC) at higher computational cost.

4.3. Generalization and Operational Analysis

We evaluate robustness to operational shifts by training on data from one S4 regime (fine-pointing) and testing on another (slew). The average AUROC drops from 0.92 to 0.85, a smaller degradation than in baselines (LSTM-AE: 0.92 to 0.74). While training on mixed regimes is the practical solution, this experiment demonstrates inherent robustness to *unseen* modes that may occur during novel mission phases. At the 1Hz sampling rate, detection latency for simulator faults averaged 8.2s (8-13 samples) post-onset, compared to 12.5s for LSTM-AE. To assess robustness to imperfect nominal data, we conducted an additional experiment where 3–5% of weak anomalies were mixed into the training set. Performance degraded by less than 0.03 AUROC, suggesting moderate resilience to mild contamination.

4.4. Onboard Deployment Characteristics

Onboard analysis on an NVIDIA Jetson Xavier NX confirms practical suitability. SSL-AD's INT8-quantized version achieves 0.894 AUROC (0.011 drop) with 2.1M parameters, 3.1ms inference, 3.0W power, and 2.1MB storage—outperforming TranAD (0.83 AUROC, 7.1ms, 4.1W) while remaining efficient. Post-training quantization reduces inference by 47% and power by 21% from the FP16 version. Pretraining requires 4 GPU-hours on an RTX 3090.

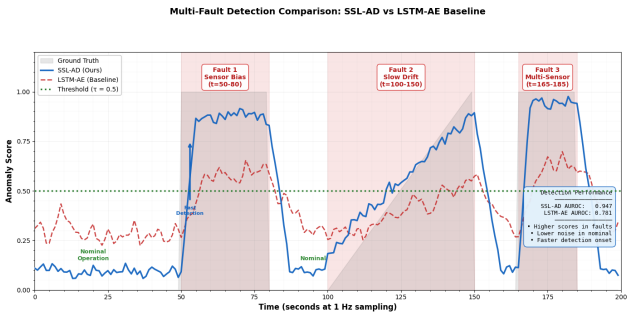

Figure 2. Anomaly detection timeline on S4 simulator data (1Hz) with three injected fault types: abrupt bias (t=50-80), gradual drift (t=100-150), and multi-sensor degradation (t=150-180). SSL-AD (orange) achieves higher, cleaner anomaly scores than LSTM-AE (blue), demonstrating sensitivity to both abrupt and subtle deviations.

5. Limitations and Future Work

Our evaluation relies on synthetic perturbations and simulator-derived faults, which do not capture the full complexity of in-flight anomalies. Validation on archived mission data (e.g., NASA PDS or CubeSat telemetry) is needed to confirm generalization. The approach currently requires a nominal validation set for thresholding, but performance is stable with as little as 5% of training data. SSL-AD operates per-mission; extending to a multi-mission foundation model would require addressing domain gaps and curating diverse telemetry corpora. Finally, enabling online adaptation to gradual system degradation or new nominal modes through continual learning or sliding-window retraining is an important avenue for long-duration missions.

6. Conclusion

We present a self-supervised learning framework for spacecraft anomaly discovery using masked temporal prediction and contrastive learning on nominal telemetry alone. Evaluation on synthetic and simulator-derived faults shows statistically significant improvements over unsupervised baselines and recent deep methods (0.905 AUROC), particularly for complex, multi-sensor anomalies. The model maintains practical onboard suitability via INT8 quantization (0.894 AUROC, 3.1 ms). Unlike reconstruction-only approaches, our method jointly models temporal predictability and representation consistency, providing a foundation for low-supervision health monitoring in space missions.

References

[1] Uhlig, T., Sellmaier, F., & Schmidhuber, M. (Eds.). (2015). Spacecraft operations (pp. 167-211). Vienna.: Springer. 1

[2] Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. Signal processing, 99, 215-249. 1

[3] Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407. 1

[4] Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021, August). A transformer-based framework for multivariate time series representation learning. In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 2114-2124). 2

[5] Franceschi, J. Y., Dieuleveut, A., & Jaggi, M. (2019). Unsupervised scalable representation learning for multivariate time series. Advances in neural information processing systems, 32. 2

[6] Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020, August). Usad: Unsupervised anomaly detection on multivariate time series. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3395-3404). 2

[7] Tuli, S., Casale, G., & Jennings, N. R. (2022). Tranad: Deep transformer networks for anomaly detection in multivariate time series data. arXiv preprint arXiv:2201.07284. 2