SOVEREIGN: To what extent does scaling the backbone size (e.g., ViT-B vs. ViT-L) in multimodal models improve robustness

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Anomaly Detection (AD) and Anomaly Localization (AL) are crucial in fields that demand high reliability, such as medical imaging and industrial monitoring. However, current AD and AL approaches are often susceptible to adversarial attacks due to limitations in training data, which typically include only normal, unlabeled samples. This study introduces PatchGuard, an adversarially robust AD and AL method that incorporates pseudo anomalies with localization masks within a Vision Transformer (ViT)-based architecture to address these vulnerabilities. We begin by examining the essential properties

1 Introduction

Analysis of: PatchGuard: Adversarially Robust Anomaly Detection and Localization through Vision Transformers and Pseudo Anomalies. Research goal: To what extent does scaling the backbone size (e.g., ViT-B vs. ViT-L) in multimodal models improve robustness to distribution shifts in brain MRI anomaly localization, as measured by Dice score degradation across different imaging protocols?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

6 papers retrieved. 10 claims extracted, 2 verified. Tribunal: 5.3/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
PatchGuard achieves competitive accuracy in non-adversarial settings.	✓	0.17
PatchGuard is evaluated under both 'Clean' and 'Adversarial' test conditions using PGD attack with $\epsilon = 8/255$.	×	0.02
PatchGuard is evaluated against the ∞PGD-1000 attack with $\epsilon = 8/255$ for detection and localization tasks.	×	0.05
Existing state-of-the-art anomaly detection and localization methods perform poorly under adversarial training condition	×	0.12
PatchGuard is trained using PGD-10 with l∞ norm under $\epsilon = 8/255$ and evaluated using PGD-1000 with l2 norms with various	×	0.02
PatchGuard demonstrates consistent robustness and effectiveness across different ϵ values as shown in Table 20.	×	0.02
In Table 20, PatchGuard achieves AD performance of 88.1 / 71.1 on MVTec AD dataset at $\epsilon = 8/255$.	×	0.05
In Table 20, PatchGuard achieves AL performance of 92.7 / 73.8 on MVTec AD dataset at $\epsilon = 8/255$.	×	0.06
Adversarial training improves model robustness in PatchGuard through a novel loss function within a ViT-based framework.	✓	0.20
The absence of anomalous samples during training prevents models from being robust against adversarial perturbations spe	×	0.06

References

- http://arxiv.org/abs/2510.02155v1
- http://arxiv.org/abs/2506.09237v2
- http://arxiv.org/abs/2111.10480v6