Towards Smarter Segmentation: Improving SAM's Anomaly Understanding with RLHF

Wan Wang, Yu-Tong Chuang, Yiu Chang, Lulin Liu

Motivation & Contribution

Motivation

Anomaly detection is essential for industrial automation but often suffers from limited high-quality segmentation data, as pixel-level annotation is costly and requires domain expertise. While Vision Foundation Models like SAM demonstrate strong generalization in segmentation tasks, they lack explicit anomaly understanding and struggle with anomaly detection.

Contribution

This project aims to investigate whether RLHF-based fine-tuning can improve SAM's anomaly segmentation performance and provide insights into the effectiveness of RLHF for this task — with the long-term goal of reducing reliance on manual annotation and enabling more automated labeling solutions.

Literature Survey

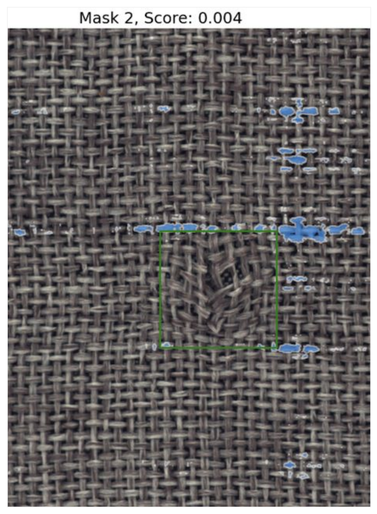
Vision Foundation Models like **SAM (Kirillov et al., 2023)** and CLIP (Radford et al., 2021) show strong generalization but struggle with anomaly segmentation due to limited domain adaptation (Jeong et al., 2023). Fine-tuning methods, including adapters (Ke et al., 2023; Wu et al., 2023) and knowledge distillation (Wang et al., 2024), improve domain alignment but often fail to capture subtle defects.

RLHF has proven effective in aligning language models with human intent (Ouyang et al., 2022) and has been explored for segmentation tasks (Huang et al., 2024). Its potential for anomaly segmentation in vision models remains largely untested — this work aims to bridge that gap.

Literature Survey

Through a Phase 1 evaluation on the MVTec AD dataset, we tested SAM's segmentation performance across 14 anomaly categories using 210 representative test images.

Despite experimenting with diverse prompting strategies, including points and bounding boxes, we observed that SAM performs poorly on anomaly segmentation. This highlights **SAM's lack of explicit anomaly understanding** — a challenge we aim to address through RLHF-based fine-tuning.



Limitation, Discussion and Future Plan

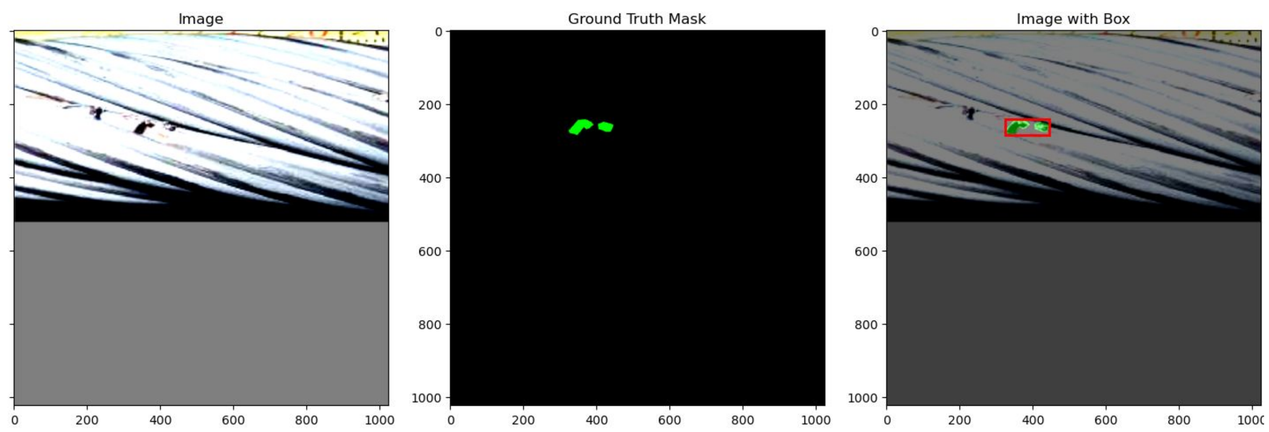
- Dataset size is limited due to data sensitivity, and we plan to explore iterative training to gradually scale up data for better performance.
- Lack of expert involvement may affect annotation quality, which we aim to address through active learning or semi-supervised methods.
- Direct fine-tuning risks catastrophic forgetting, and future work will consider adapter-based tuning or RLHF as alternative strategies.

Proposed Ideas

Phase 1 - "Cold Start": Conduct literature review, select SAM as the baseline model, and evaluate its anomaly segmentation performance on the MVTec AD dataset. Identify SAM's limitations across different prompting strategies and anomaly categories.

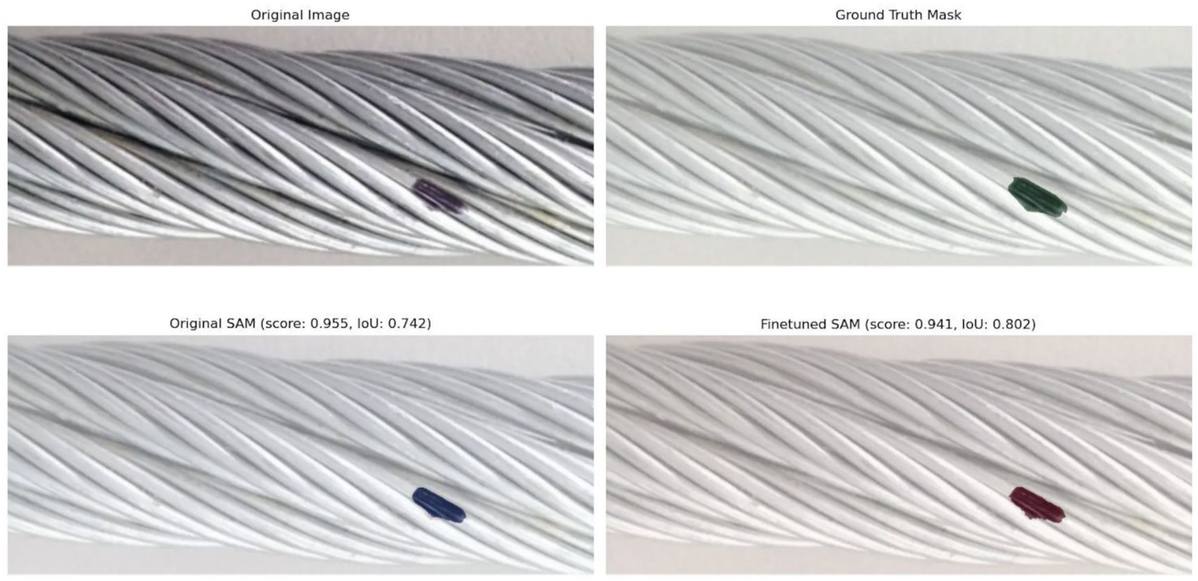
Phase 2 - SFT: 1) Developed training and inference pipeline and 2) Train a baseline model with pixel-level annotations to establish a reference for comparison.

Phase 3 - Introduce RLHF: Introduce Reinforcement Learning with Human Feedback (RLHF) into the fine-tuning process. Use human-in-the-loop reward signals to guide SAM toward improved segmentation quality on anomaly detection tasks.



An example of one sample of the Vision Dataset. VISION Workshop. VISION-Datasets: Industrial Anomaly Segmentation Dataset. Available at:https://huggingface.co/datasets/VISION-Workshop/VISION-Datasets (accessed April 2025).

Current Results, Comparison and Main Findings



Current Results

We evaluated SAM's segmentation performance on anomaly detection tasks using steel cable defect samples. Above is one of the test results.

Comparison

Quantitative evaluation shows consistent improvement across key metrics after fine-tuning. The fine-tuned SAM outperforms the original SAM by ~1.32% in IoU, with noticeable gains in Dice score and confidence. However, the overall improvement remains limited due to the small size of the fine-tuning dataset.

Main Findings

- Fine-tuning SAM on domain-specific anomaly data improves segmentation accuracy, especially in critical defect regions.
- The performance gain is measurable but modest, suggesting that data scale is a key factor for effective fine-tuning.