

Research and Application of Parameter-Efficient Adaptation (LoRA) on Vision-Language Models for Industrial Anomaly Detection

Ho Dang Thanh Ho | hohdt.20@grad.uit.edu.vn | University of Information Technology

THE CHALLENGE IN INDUSTRIAL QUALITY CONTROL

Traditional AI is Inflexible and Expensive

Standard models (CNNs, Autoencoders) need extensive, labeled defect data and must be retrained from scratch for new products, increasing time and cost.

Foundation Models Face a "Domain Gap"

A model like CLIP, trained on web images, understands "a bottle" but fails to recognize a "bottle with a microscopic scratch," which is critical in industrial settings.

Full Fine-Tuning is Impractical

Retraining the entire CLIP model (be of millions of parameters) requires powerful GPUs and risks catastrophic forgetting, where the model loses its original powerful knowledge.

How can we adapt CLIP for Industrial tasks efficiently?

The goal is to achieve high accuracy in a low-data (few-shot) environment while minimizing computational costs.



EXPERIMENTAL FRAMEWORK

Dataset: MVTec AD

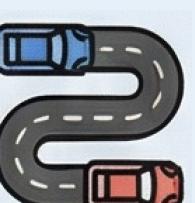
A comprehensive, real world benchmark dataset for unsupervised anomaly detection, featuring 15 different industrial object categories.

Evaluation Metric: AUROC

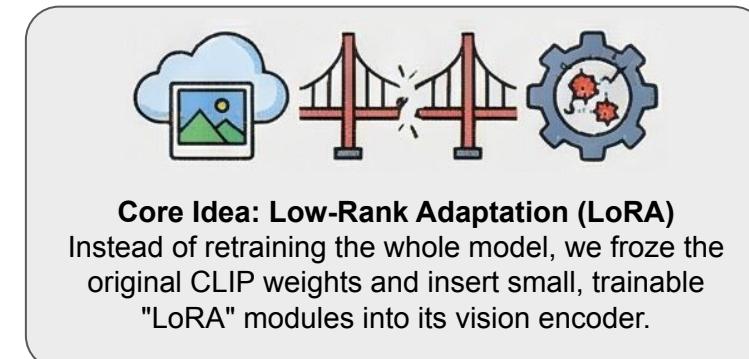
The model's performance will be measured by the Area Under the Receiver Operating Characteristic (AUROC) curve, a standard for classification tasks.

Benchmarking Against State-of-the-Art:

The proposed method results will be compared against leading models in the field, including WinCLIP and AnomalyCLIP.



PROPOSED SOLUTION: AN EFFICIENT TWO-STREAM ADAPTATION



1. Image Encoder (Vision Stream)

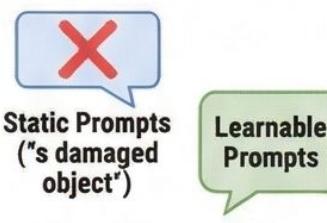


CLIP Vision Transformer (ViT) - Frozen

Trainable LoRA Modules

Image Features

2. Text Encoder (Language Stream)



We replace static text prompts with "Learnable Prompts," allowing the model to automatically find the optimal vector representations for "Normal" and "Anomaly".

CLIP Text Encoder - Frozen

Normal Prompt Representation Anomaly Prompt Representation

3. Few-Shot Training

The model is trained using only a handful of "normal" images (e.g. 4, 8, or 16 shots). The training process optimizes only the LoRA and prompt parameters.

4. Anomaly Detection

During inference, the system calculates the cosine similarity between a new product's image features and the learned "Normal" prompts to generate an anomaly score.

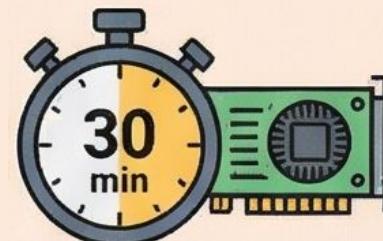
Anomaly Score
(Normal vs. Anomaly)

THE CHALLENGE IN INDUSTRIAL QUALITY CONTROL



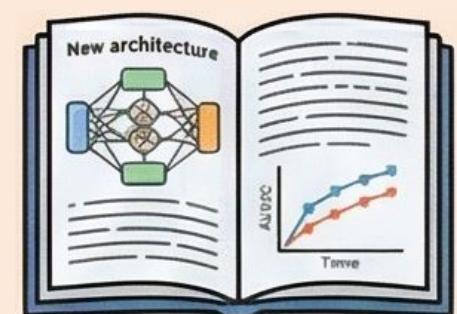
High Accuracy: AUROC > 90%

The model is expected to achieve an image-level AUROC score exceeding 90% on the MVTec AD dataset.



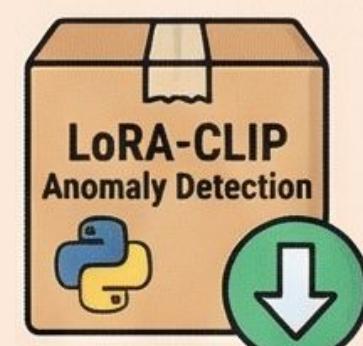
Proven Efficiency

The method will be validated to run on consumer-grade GPUs (like NVIDIA T4 or RTX 3060) with a training time of less than 30 minutes per product class.



Scientific Contribution

The research will deliver a novel, effective architecture combining CLIP and LoRA, along with an ablation study analyzing the impact of shot count on accuracy.



Scientific Contribution

The research will deliver a novel, effective architecture combining CLIP and LoRA, along with an ablation study analyzing the impact of shot count on accuracy.