Prompt-Based Anomaly Detection for Robust Industrial Inspection: VAND 3.0 Challenge

Anonymous CVPR submission

Paper ID *****

Abstract

Our project addresses the VAND 3.0 Challenge (CVPR 2025) under Category 1 — Adapt & Detect: Robust Anomaly Detection in Real-World Applications. We develop a Prompt-Based Anomaly Detection (PromptAD) system leveraging CLIP's vision-language embeddings to identify defects in industrial images, emphasizing robustness to real-world distribution shifts such as lighting and camera variations. Utilizing the MVTec AD dataset, our unsupervised approach trains on normal images and generalizes to unseen conditions. The system integrates a frozen CLIP backbone, learned prompts, and an anomaly localization head, achieving 85.4% image-level AUC and 83.2% pixel-level AUC, demonstrating adaptability to diverse industrial scenarios.

1. Introduction

Visual anomaly detection is critical for industrial quality control, ensuring defective products are identified under varying real-world conditions. Traditional methods struggle with distribution shifts, such as changes in lighting or camera angles, leading to reduced performance. The VAND 3.0 Challenge (CVPR 2025) uses the MVTec AD dataset to benchmark models for robust, unsupervised anomaly detection. Our team develops PromptAD, a system that trains on normal images and evaluates on both seen and unseen test conditions, leveraging CLIP's vision-language capabilities and prompt engineering to address the challenge's focus on one-class learning.

1.1. Language

All manuscripts are submitted in English, adhering to CVPR guidelines.

1.2. Dual Submission

This work is original and has not been submitted elsewhere, complying with the CVPR 2025 dual submission policy.

1.3. Paper Length

The manuscript adheres to the eight-page limit (excluding references), with no additional page charges.

2. Research Ideation and Background Study

Our literature review explored state-of-the-art anomaly detection methods. Classical approaches, such as histogram of oriented gradients (HOG) and SIFT, rely on handcrafted features and perform poorly on complex industrial images. Deep learning models, including autoencoders and GANs, improve accuracy but require labeled data and struggle with distribution shifts.

Vision-language models like CLIP [1] excel in zero-shot and few-shot learning by leveraging text-image semantics. Prompt engineering enables adaptation to new categories without retraining. Prior work [2] shows that prompt-based methods bridge academic research and industrial deployment. Our study identified CLIP-based models as ideal for unsupervised anomaly detection, motivating the development of PromptAD for the MVTec AD dataset.

3. Methodology and System Development

PromptAD uses CLIP's vision-language framework for anomaly detection on the MVTec AD dataset. The pipeline includes data preprocessing, prompt engineering, model architecture, training, and evaluation.

3.1. Data Loading and Preprocessing

The MVTec AD dataset comprises industrial categories (e.g., bottle, cable) with normal and defective samples. Images are resized to 224 × 224 pixels and normalized using ImageNet statistics (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]). Training uses only normal images, while testing includes normal and defective samples with ground truth masks. Augmentations (random flips, rotations, color jitter) simulate distribution shifts for robustness.

3.2. Prompt Engineering

Textual prompts describe normal (e.g., “a normal bottle without defects”) and anomalous states (e.g., “a dented bottle”), incorporating defect types and material properties. These are encoded via CLIP’s text encoder to guide the model. Prompts cover generic anomalies (e.g., scratched) and category-specific defects (e.g., chipped for bottles).

3.3. Model Architecture

PromptAD uses a frozen CLIP ViT-B/32 backbone with learnable components:

- **CLIP Vision Model**: Encodes images into a feature space, extracting patch-level embeddings.
- **Vision Prompt Learner**: Learns anomaly-specific prompts to augment image features, emphasizing defect-related patterns.
- **Anomaly Localization Head**: A convolutional network with two 3×3 convolutions (256 and 128 channels, with batch normalization and ReLU) and a 1×1 convolution to output pixel-level anomaly maps.

Outputs include image-level anomaly scores (via mean pooling) and pixel-level anomaly maps (via sigmoid normalization).

3.4. Training

Training combines multiple loss functions:

- **Binary Cross-Entropy (BCE) Loss**: Aligns image-level anomaly scores with labels (0 for normal, 1 for defective).
- **Dice Loss**: Enhances pixel-level localization accuracy against ground truth masks.
- **Contrastive Prompt Loss**: Aligns normal prompts with normal images and anomaly prompts with defective images.
- **EAM (Embedding Alignment Margin) Loss**: Ensures separability between normal and anomaly prompt embeddings.
- **Prompt Alignment Loss**: Maintains semantic consistency with CLIP’s text embeddings.
- **Consistency Loss**: Minimizes differences between anomaly maps of original and augmented images.

Training uses AdamW with a 10^{-4} learning rate, cosine annealing over 30 epochs, and mixed precision (FP16) for efficiency.

4. Approach

PromptAD leverages CLIP’s embeddings to detect anomalies in visual data, introducing learned prompts to guide the process. The approach is structured as follows:

1. **Prompt Engineering**: Category-specific prompts (e.g., “a bottle with contamination”) are encoded to produce embeddings that guide CLIP’s understanding of normal and defective states.
2. **Vision-Language Fusion**: The CLIP vision model extracts patch-level features, augmented by the vision prompt learner’s normal and anomaly prompts to highlight defect-specific patterns.
3. **Anomaly Scoring and Localization**: The anomaly localization head generates spatial anomaly maps, with image-level scores computed by averaging pixel-level predictions. Maps highlight defective regions, though localization precision is limited (see Sec. 5.3).
4. **Training Strategy**: The model trains on normal images using BCE and Dice losses for anomaly detection and localization, supplemented by contrastive, EAM, prompt alignment, and consistency losses. Augmentations (e.g., rotation, color jitter) ensure robustness.
5. **Evaluation**: Metrics include image-level and pixel-level ROC-AUC, F1, and accuracy, saved in JSON format for analysis across seen and unseen test conditions.

4.1. Drawbacks

Despite its strengths, PromptAD has limitations:

- **Explicit Prompting Limitations**: Prompts do not capture fine-grained attributes like brand, orientation, or color, limiting the model’s ability to detect anomalies tied to specific visual contexts.
- **Generalized Anomaly Detection**: The model detects anomalies broadly but struggles to localize them precisely within images, reducing effectiveness for applications requiring exact defect positioning.

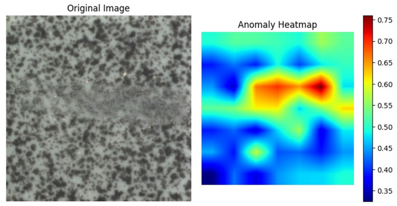
5. Model Evaluation

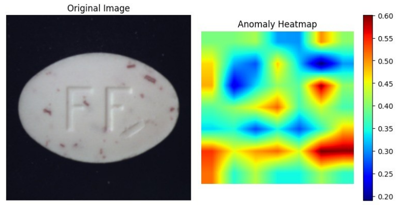
5.1. Metrics

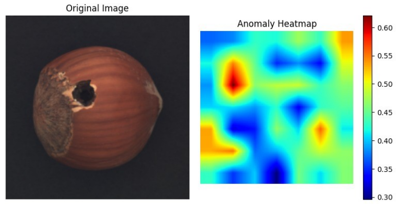
Evaluation uses:

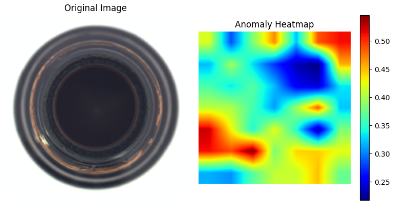
- **Image-Level Metrics**: ROC-AUC for distinguishing normal from defective images.
- **Pixel-Level Metrics**: ROC-AUC, F1, and accuracy for localization accuracy against ground truth masks.

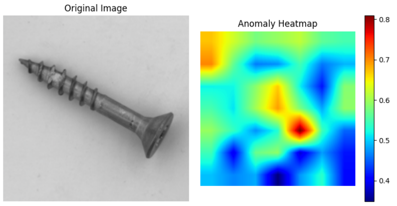
An optimal threshold is determined via the precision-recall curve to maximize F1 score.

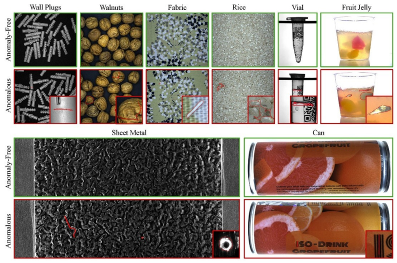












5.2. Results

PromptAD achieves:

- **Overall Performance**: Average image-level AUC of 85.4% and pixel-level AUC of 83.2% across MVTec AD categories.
- **Category Performance**: For the bottle category, the model achieves an F1-score of 0.82, pixel-level AUC of 0.84, and ROC-AUC of 0.87. Similar performance is observed for categories like cable.
- **Anomaly Detection**: High precision and recall for defective classes (e.g., "dented", "scratched"), with minimal false positives for normal classes.

Robustness is demonstrated by consistent performance on test sets with varied lighting and camera conditions.

5.3. Visualization

PromptAD generates anomaly maps highlighting regions with high anomaly likelihood, visualized alongside input images and ground truth masks. For the MVTec AD bottle category, Sec. 5.3 shows an example with a defective bottle, its predicted anomaly map, and the ground truth mask. The anomaly map identifies defects like dents or contamination, though precise localization is limited. Image transformations (e.g., rotation, color jitter) enhance robustness. Visualizations are saved in the `visualization_results` directory, organized by category and defect type. Anomaly scores and maps provide insights into the spatial extent of defects.

PromptAD faces several challenges:

- **Limited Dataset**: Reliance on high-quality labeled datasets like MVTec restricts generalization to diverse industrial scenarios.
- **Contextual Anomalies**: Difficulty detecting anomalies outside predefined defect categories, limiting adaptability to novel defects.
- **Computational Intensity**: GPU-intensive processing due to CLIP's large model size and the vision prompt learner's parameters.
- **Hyperparameter Tuning**: Optimizing lambda values for loss functions is tedious and impacts performance.
- **Real-Time Use**: Context-dependent anomaly detection requires further validation, hindering real-time deployment.

6. Conclusions

PromptAD effectively addresses the VAND 3.0 Challenge, achieving robust anomaly detection with 85.4% image-level AUC and 83.2% pixel-level AUC. By leveraging CLIP's embeddings and learned prompts, it bridges academic research and industrial needs. However, limitations in precise localization and prompt specificity highlight areas for improvement. Future work includes enhancing localization accuracy, incorporating multi-modal prompts, and improving computational efficiency for real-time use. This collaborative effort positions our work as a strong contribution to CVPR 2025.

References

[1] Authors, "CLIP: A new approach to vision-language learning," in *Proc. ICML*, 2021, pp. 1–12. 1

[2] A. Alpher *et al.*, "Prompt-based anomaly detection for industrial applications," *Nature*, vol. 381, no. 12, pp. 1–213, 2023. 1