# PromptAD: Zero-shot Anomaly Detection using Text Prompts

Yiting Li[1], Adam Goodge David[1], Fayao Liu[1] *, Chuan-Sheng Foo[1 2] *
[1] Institute for Infocomm Research (I$^2$R), A*STAR, Singapore
[2] Centre for Frontier AI Research (CFAR), A*STAR, Singapore
{li_yiting, Goodge_Adam_David, Liu_Fayao, foo_chuan_sheng}@i2r.a-star.edu.sg

## Abstract

*We consider the problem of zero-shot anomaly detection in which a model is pre-trained to detect anomalies in images belonging to seen classes, and expected to detect anomalies from unseen classes at test time. State-of-the-art anomaly detection (AD) methods can often achieve exceptional results when training images are abundant, but they catastrophically fail in zero-shot scenarios with a lack of real examples. However, with the emergence of multi-modal models such as CLIP, it is possible to use knowledge from other modalities (e.g. text) to compensate for the lack of visual information and improve AD performance. In this work, we propose PromptAD, a dual-branch framework which uses prior knowledge about both normal and abnormal behaviours in the form of text prompts to detect anomalies even in unseen classes. More specifically, it uses CLIP as a backbone encoder network and an additional dual-branch vision-language decoding network for both normality and abnormality information. The normality branch establishes a profile of normality, while the abnormality branch models anomalous behaviors, guided by natural language text prompts. As the two branches capture complementary information or 'views', we propose a 'cross-view contrastive learning' (CCL) component which regularizes each view with additional reference information from the other view. We further propose a cross-view mutual interaction (CMI) strategy to promote the mutual exploration of useful knowledge from each branch. We show that PromptAD outperforms existing baselines in zero-shot anomaly detection on key benchmark datasets and analyse the role of each component in ablation studies.*

## 1. Introduction

Anomaly detection (AD) is important in a wide range of applications, such as industrial product inspection, network security and autonomous driving. Although existing anomaly detection techniques have achieved impressive performance in many cases, they are only evaluated on classes of data that the model has observed during training. In real-world scenarios, a model is often required to evaluate samples in an open-world setting: that is, to detect anomalies in previously unseen and novel classes. Despite its importance, this cross-category generalization has been largely overlooked in existing literature. Thus, in this work, we target this problem of **zero-shot anomaly detection (ZSAD)** (Fig. 1), where the model is trained to detect anomalies in the "*hazelnut*" and "*metal nut*" classes and expected to generalise to detect anomalies in the previously unseen "*tile*" and "*capsule*" classes without further training.

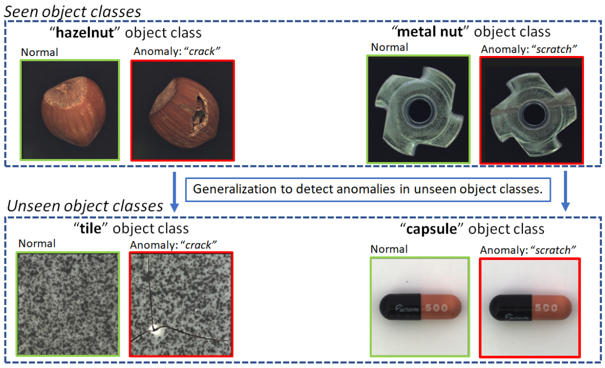


Figure 1. The ZSAD problem addressed in this work. The model is trained to detect anomaly types like "crack" and "scratch" in the set of known object classes "hazelnut" and "metal nut". At test-time, the model is tasked with detecting similar anomaly types in previously unseen object classes "tile" and "capsule".

ZSAD is a difficult task in the absence of real data from the unseen classes. However, prior information about the expected behaviour of normal data, as well as the potential behaviour of anomalies, is often known even without visual examples [22, 23]. For example, quality control engineers can describe how a manufactured product should and should not appear in the form of natural language (text). With

*indicates corresponding authors. This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds (Grant No. A20H6b0151).

the emergence of vision-language models like CLIP [18], which has demonstrated its capability for image-level zero-shot classification, this natural language information can be encoded to obtain semantic representations of both normality as well as abnormality, which can compensate the lack of visual examples and help in the anomaly detection task.

To this end, we propose an effective and flexible framework named PromptAD, which equips the model with both a normality view and abnormality view that leverage rich information from the language mode (text) to reduce dependence on image data. Information sharing between the two views further refines the representation learning and improves generalization. PromptAD is built on the CLIP model, which is used to encode data from both image and text modalities. On top of this backbone, PromptAD decodes these representations into feature maps for anomaly detection through both 'intra-view' and 'cross-view' modeling. Intra-view modeling captures knowledge specific to each view through two parallel vision-language decoding networks (one each for normality and abnormality). Cross-view modeling shares information between the two views via a cross-view contrastive learning strategy (CCL), which regularizes the intra-view training with additional reference information from the other complementary view, and a cross-view mutual interaction (CMI) strategy is proposed to facilitate the explicit knowledge transfer from each other.

Existing uni-modal AD methods rely exclusively on visual information, which hinders their generalization beyond the base training classes. Moreover, uni-modal training often utilizes a considerable amount of training data, which incurs extra data collection cost. In contrast, PromptAD has two main advantages: strong transferability and high data-efficiency. It utilizes semantic knowledge from text prompts to learn representations that are more reliable and more transferable to new classes. We further observe that PromptAD requires much less training data to achieve comparable performance to methods trained on a large amount of data. This is because a well-designed text prompt can contain rich semantic information that effectively distills the information contained in a large number of images. We demonstrate these advantages of PromptAD through extensive experiments. In summary, the contributions of this paper are as follows:

1. We propose PromptAD which efficiently adapt the pre-trained CLIP features aligned with language for zero-shot anomaly detection. The proposed approach enables training one unified and generalizable model without any fine-tuning when adapting to new classes.

2. Our framework effectively aggregate the semantic knowledge from both normality view and the abnormality view. A Cross-View Contrastive Learning (CCL) strategy is proposed to readjust the optimization difficulty of the intra-view modeling, and a Cross-View Mutual Interaction (CMI) approach is proposed to promote the explicit knowledge transfer between two complementary branches.

3. We show that our approach significantly improves the anomaly detection performance over existing methods on widely used AD benchmarks.

## 2. Related Works

Existing methods can be divided into unsupervised and supervised methods. Unsupervised methods model the normal sample distribution and exclude anomalies in training. Reconstruction-based methods [2, 6, 19, 26] generate reconstructed images and then use reconstruction errors between the input image and its reconstruction to detect anomalies. GAN-based models detect anomalies based on the ability of the generator to generate a given test sample [1, 20, 24]. In contrast, supervised methods find a better decision boundary between normal and anomalous samples by leveraging synthetically generated anomalies [13] or a small number of real anomalies. DevNet [17] encourages the anomaly score of normal samples towards a common center, whereas MLEP [14] maximises the pair-wise distances between normal and anomalous features. DRA [7] uses a multi-head neural network to learn disentangled representations for different types of anomalies separately.

As mentioned, these methods rely on the availability of training data from all classes and are ill-equipped to detect anomalies from previously unseen classes at test time. On the other hand, PromptAD can detect anomalous samples from novel classes not observed during training; it achieves this by incorporating additional domain knowledge from natural language descriptions of normal and abnormal behaviour to compensate for the lack of real samples.

It worth to take note that WinCLIP [10] also uses CLIP for zero-shot anomaly detection. However, it uses representations of query images directly from the fixed CLIP model parameters without further training. This may be suitable for domains which are closely related to the original data used to train CLIP, however, as noted in its original paper [18], CLIP under-performs when applied to domains that are not well represented in this training distribution. In contrast, PromptAD builds two trainable branches on top of a frozen CLIP model, which combines CLIP's multi-modal capabilities with task-specific representation learning and directly optimises the model for the anomaly detection task at hand. Therefore, PromptAD can effectively ameliorate the distribution shift issue and thus achieves better pixel-level AD performance than WinCLIP (Table 2).

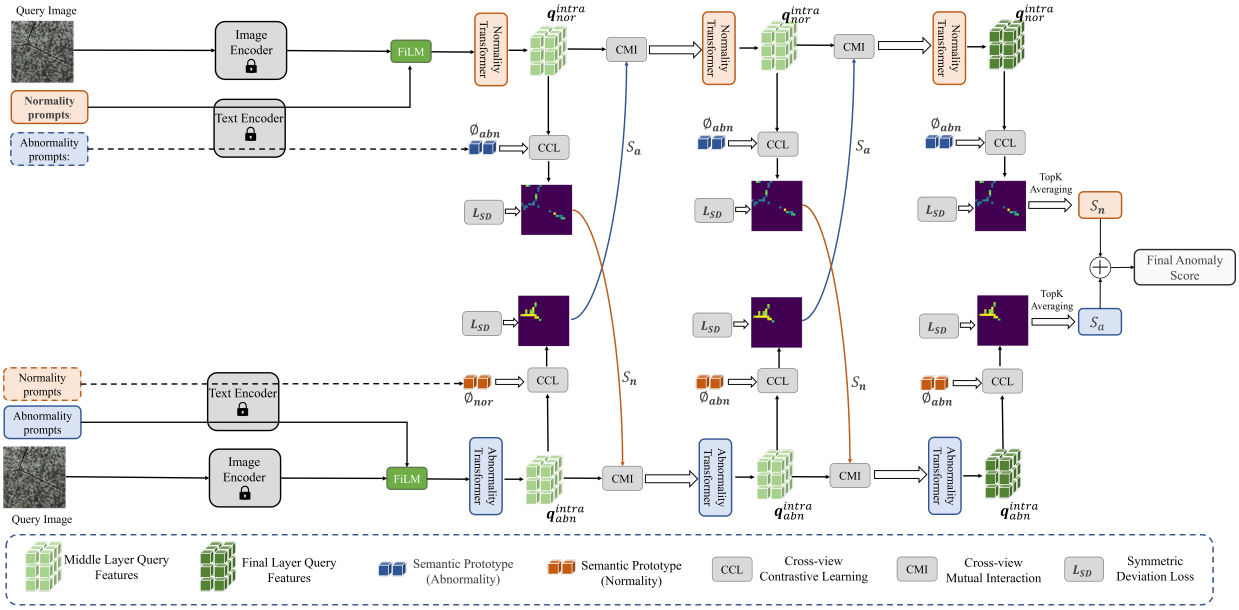


Figure 2. An overview of our PromptAD framework. It uses a dual-branch network design, which aims to detect anomalies from both normality (upper part of the figure) and abnormality (lower part of the figure) views. The normality branch learns to detect out-of-distribution patterns by modeling the normal data conditioned on images and normality text prompts. The abnormal branch attempts to directly identify anomalies by conditioning on images and abnormality text prompts. A cross-view contrastive learning (CCL, details in Fig. 3 and Sec. 3.4) approach is proposed to incorporate complementary information from the opposite view for better anomaly targeting in each branch. The two branches further explore knowledge from each other through cross-view mutual interaction (CMI, Sec. 3.5).

# 3. Anomaly Detection with Anomaly-Aware Text Prompts

## 3.1. Problem Description

We focus on the ZSAD problem formulated as follows. We train a model on the base data, consisting of $N$ normal samples and a few anomalous samples from a base (or seen) class $\mathcal{C}_b$. We then test the model on a set of novel classes, $\mathcal{C}_u$, without additional training to evaluate zero-shot performance, where $\mathcal{C}_b \cap \mathcal{C}_u = \emptyset$. Additionally, for each class in $\mathcal{C} = \mathcal{C}_b \cup \mathcal{C}_u$, we define two types of semantic knowledge in the form of text prompts, where the normality prompt $P_{nor}$ describes the visual appearance of normal patterns, and the abnormality prompt $P_{abn}$ describes typical anomaly appearances (details on how to construct text prompts are given in Sec. 4.1). Our goal is to train an anomaly detector $f : (X, P_{nor}, P_{abn}) \rightarrow R$ from a single base class $\mathcal{C}_b$ to detect anomalies on unseen classes $\mathcal{C}_u$, by assigning larger scores to anomalies than normal samples.

## 3.2. PromptAD Framework

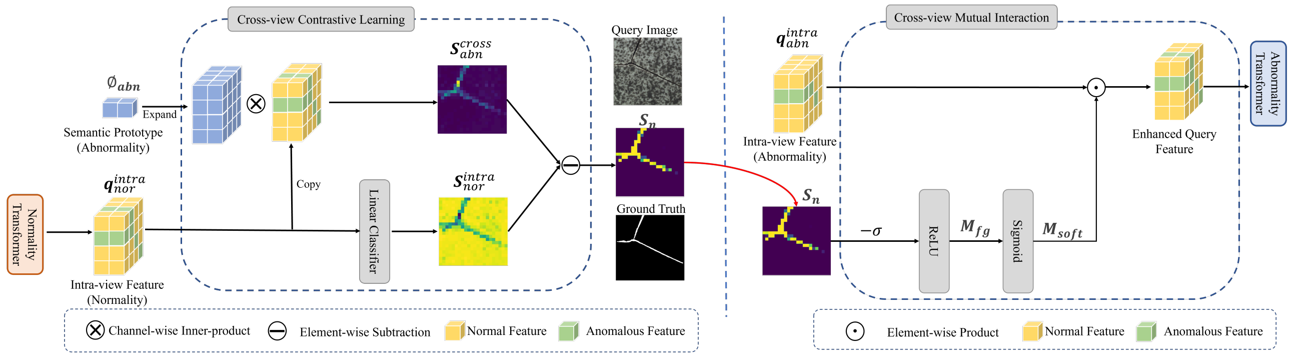
Here we describe the PromptAD framework (Fig. 2), giving an overview before detailing the individual components. PromptAD has two complementary branches built on top of a shared CLIP model. The abnormality branch directly models the distribution of the available anomaly samples (Sec. 3.3.1) while the normality branch models the distribution of the normal samples, directly measuring the conformity of a query image to the normality descriptions (Sec. 3.3.2). To promote complementary knowledge transfer between the two views, a cross-view contrastive learning (CCL, Sec. 3.4) approach is proposed to regularize intra-view training with additional reference information from the complementary view. Furthermore, the two branches also explicitly explore knowledge from each other through cross-view mutual interaction (CMI, Sec. 3.5).

## 3.3. Intra-view Modeling

### 3.3.1 Abnormality Branch

The abnormality branch (lower part of Fig. 2) learns to detect anomalies according to information provided by the abnormality text prompts. In this branch, we obtain the visual embedding of a given image $x$ from the CLIP image encoder and the semantic embedding of the abnormality prompt $P_{abn}$ from the CLIP text encoder. We then fuse these embeddings using FiLM [8] as in [15].

The fused feature embeddings are fed as input to a transformer-based decoder to generate output features $q_{abn}^{intra} = \{q_{abn}^1, q_{abn}^2, \cdots, q_{abn}^L\}$, where each $q_{abn}^i \in R^D$ corresponds to a local area of the input image and $D$ is the



(a) (b)

Figure 3. **(a)**. An illustration of the proposed Cross-view Contrastive Learning (CCL) pipeline. Using the normality branch as an example, to generate the intra-view normality score, a linear classifier is applied on each spatial location of $q_{nor}^{intra}$ for point-wise normality score predictions. To generate the cross-view abnormality score, the normality-aware features $q_{nor}^{intra}$ are then compared with the semantic prototype of the abnormality view $\phi_{abn}$ with inner-product similarity. Finally, the two-view scores are fused together through element-wise subtraction for unbiased anomaly detection $s_n$. **(b)**. An illustration of the proposed Cross-view Mutual Interaction (CMI).The detection score map $s_n$ from the normality branch are then used as an auxiliary attention to discover those anomalies that are neglected by the abnormality branch.

channel size of the output feature space. A linear classifier (1x1 convolution) is then applied to each $q_{abn}^{i} \in q_{abn}^{intra}$ to output an anomaly score for each spatial location. We denote the obtained 2D score map as $s_{abn}^{intra}$, which is referred as the **Intra-view Abnormality Score**. Regions that contain anomalies recorded by the abnormality prompt $P_{abn}$ should produce higher responses than normal regions.

### 3.3.2 Normality Intra-view Modeling

The normality branch (upper part of Fig. 2) is a similar decoder network, but it is instead designed to directly measure the conformity of a query image to the normal profile defined by the normality prompts, in an adjacent process to the abnormality branch. As the abnormality prompts cannot describe every possible type of anomaly (which could be infinite), modeling normality is also important to detect anomalies as those samples that do not conform with the normal profile. Similarly, the produced prediction $s_{nor}^{intra}$ is referred as **Intra-view Normality Score**.

## 3.4. Cross-view Contrastive Learning

As an anomaly usually refers an irregularity or deviation from the standard pattern, the single-view approach (intra-view modeling) may not be optimal for detecting certain anomalies that do not exhibit obvious irregularities. For example, in the popular MVTec dataset [3], the "*transistor*" class has a type of anomalies named "misplaced", where a transistor in good condition is shifted away from the right location. Such anomalies can only be effectively detected given a reference standard, e.g.,"a normal transistor should be placed vertically in the middle line of the circuit board". To better recognize such anomalies, we propose a cross-view contrastive learning (CCL) approach (illustrated in Fig. 3) which enhances the intra-view modeling with additional reference information from the opposing view. In particular, CCL learns anomaly scores based on the residuals between the two complementary views in a learned feature space. This process will now be detailed for each branch.

### 3.4.1 Abnormality Branch

The abnormality branch is regularized with complementary information from the normality view. After extracting the semantic embedding of the normality prompt from the CLIP text encoder, we first reduce its channel dimension with a 1x1 convolution layer to match with the output-space dimension. The obtained semantic feature is regarded as the normality prototype $\phi_{nor}$. Given the intra-view abnormality features $q_{abn}^{intra}$ extracted from a query image, we also compute a **Cross-view Normality Score** $s_{nor}^{cross}$ to measure the closeness between the normality prototype $\phi_{nor}$ and $q_{abn}^{intra}$. In particular, $s_{nor}^{cross}$ is a 2D score map with the same size as the intra-view abnormality score $s_{abn}^{intra}$, which is obtained by measuring inner-product similarity between $\phi_{nor}$ and the feature vector of each spatial location in $q_{abn}^{intra}$, where $s_{nor}^{cross} = \phi_{nor} \cdot q_{abn}^{intra}$ ($\cdot$ denotes inner product).

Oppositely from $s_{abn}^{intra}$, images with high responses in $s_{nor}^{cross}$ are those that have higher possibilities of being normal. The final prediction $s_a$ of the abnormality branch is given by an element-wise subtraction of the two scores:

$$s_a = s_{abn}^{intra} - s_{nor}^{cross}. \quad (1)$$

In doing so, our model learns generalized and anomaly-aware representations rather than over-fitting to the limited anomaly modes described in abnormality prompts.

### 3.4.2 Normality Branch

Anomalies often share many compositional patterns with normal samples and their anomalousness may be very subtle, for example "an anomaly glass bottle with small crack" vs. "a normal sample with small scratch within acceptable limits" is difficult to distinguish. As the intra-view learning of the normality branch focuses on the normality reference, it struggles to detect such anomalies; carefully-defined abnormality prompts are useful for defining this boundary.

For this reason, we similarly obtain the abnormality semantic prototype $\phi_{abn}$ to highlight those hard anomalies. Formally, given the extracted intra-view features $q_{nor}^{intra}$, we compute its inner-product similarity with $\phi_{abn}$ to obtain the **Cross-view Abnormality Score** $s_{abn}^{cross}$, where $s_{abn}^{cross} = \phi_{abn} \cdot q_{nor}^{intra}$. Higher values in $s_{abn}^{cross}$ indicate that the corresponding locations in $q_{nor}^{intra}$ are more similar to the abnormality prototype $\phi_{abn}$, and are more likely to be anomalous. On the other hand, the normal regions in the query features $q_{nor}^{intra}$ can hardly match the abnormality features in $\phi_{abn}$ with high similarity, leading to lower values in $s_{abn}^{cross}$. To this end, the final anomaly score map of normality branch is obtained by:

$$s_n = s_{abn}^{cross} - s_{nor}^{intra}. \tag{2}$$

## 3.5. Cross-view Mutual Interaction

The two branches capture complementary information: the abnormality branch is more effective for anomalies that are well defined by the abnormality prompts while the normality branch is effective for more general anomalies that are missed by the abnormality prompts; we believe that the attention maps[1] learned by one branch are helpful in discovering overlooked anomalies by the other [27]. Therefore, rather than separate training, the two branches can each benefit through mutual interaction [16, 28]. In particular, attention maps from the intermediate layers can highlight those regions in the image that were important for the network's decisions [11, 12]. Hence, we propose to utilize the intermediate attentions of one branch to extend possible anomalous regions for the other branch, so that the other branch can better detect anomalies from regions that it has missed.

For example, the detection score map $s_n$ output by the normality branch can serve as a soft attention map $M_{soft}$ for guiding the abnormality branch to discover hard anomalies that are neglected by itself. In particular, the intermediate feature map $q_{abn}^{intra}$ of the abnormality branch is refined through spatial-wise multiplication with the attention map $M_{soft}$. Such auxiliary attention then serves as feature selectors to highlight anomalous regions on query features during the forward pass, as well as gradient selectors to correct such false-negative errors during the backward pass.

To ensure that the attention maps generated from middle layers can cover most of the anomalous regions, we additionally train the intermediate layers with supervision signal, i.e., the symmetric deviation loss detailed in Sec 3.6. However, the obtained raw score map $s_n$ could be noisy due to its unbounded prediction values and diverse backgrounds. To highlight the true anomaly regions, regions with activation values less than $\sigma$ in $s_n$ will all be treated as the background zone and masked out, such that $s_n$ is tailored towards the real anomalies. Therefore, we define the remodeled foreground attention map $M_{fg}$ as:

$$M_{fg} = \mathsf{ReLU}(s_n - \sigma). \tag{3}$$

Here we use $\sigma = 0.5$ for all experiments. The ReLU function ensures that only highly scoring regions (greater than a threshold $\sigma$) are activated. We observed that the resulting candidate regions cover most of the real anomalous regions for the vast majority of inputs. Then $M_{fg}$ is then used to generate a soft mask via a sigmoid function:

$$M_{soft} = \mathsf{sigmoid}(M_{fg}). \tag{4}$$

Next, $M_{soft}$ obtained from the normality branch is used as an attention map to re-weight the intra-view features $q_{abn}^{intra}$ of the abnormality branch, so that the neglected anomalies can be highlighted (Fig. 3). This drives the two branches into a mutual-guidance state so that anomalies missed by one branch can be effectively captured through the complementary attentions from the other branch.

## 3.6. Training and Inference

**Symmetric Deviation Loss** The goal of training is to enforce statistically significant deviations between the scores of anomalies from those of normal samples. Inspired by the success of the deviation loss in [17], we use a predefined Gaussian distribution to generate a set of reference scores $R$, which serves as an anchor to guide the learning of anomaly scores: $dev(x) = \frac{S-\mu_R}{\delta_R}$, where $\mu_R$ is the mean and $\delta_R$ the standard deviation of scores in $R$. We then introduce a symmetric deviation loss to enlarge the gap between intra-view and cross-view scores:

$$\begin{aligned} L(x) = (1-y) \cdot max(0, \alpha + dev(x)) \\ +y \cdot max(0, \alpha - dev(x)), \end{aligned} \tag{5}$$

where $y = 1$ if $x$ is an anomaly and $y = 0$ if $x$ is normal. $\alpha$ is a hyper-parameter defining the margin between the scores of normal and anomalous samples. The mean value of the reference score set serves as the classification boundary. The proposed loss enforces a positive deviation

[1] Attention maps here refer to the generated anomaly score maps [21]. We will use these two terms interchangeably.

of at least $\alpha$ between the classification boundary $\mu$ and the anomaly scores of anomalies in the upper tail, while enforcing a negative deviation of at least $-\alpha$ between the classification boundary $\mu$ and the anomaly scores of normal samples in the lower tail. In our experiments, we follow previous works and set $\mu_R$ = 0, $\delta_R = 1$ and $\alpha = 5$. Importantly, the symmetric deviation loss is applied and back-propagated through multiple layers for the anomaly scores calculated from the abnormality branch in Eq. 1 and the normality branch in Eq. 2 separately.

**Inference** To perform unbiased anomaly classification, we combine the scores from the normality branch $s_n$ and abnormality branch $s_a$ as they contain complementary information. The first learns normality representation that enables the universal anomaly detection on unseen categories. The latter incorporates more specific knowledge of abnormality retrieved via language, which helps learn discriminative features for the detection of hard anomalies (e.g., anomalies similar to normal samples). We thus compute the final anomaly score as $s = s_n + s_a$.

## 4. Experiments

### 4.1. Experiment Settings

We experiment on two challenging real-world benchmark datasets for industrial anomaly detection which are MVTec AD and AITEX. MVTec AD consists of 4096 normal and 1258 anomalous images split across 15 object categories, while AITEX contains 7 types of fabric defects. Experiments were conducted using the leave-one-category-out setting [9], i.e., a target category was chosen to be tested, while other categories in the dataset are used for training. As labelled anomalies are difficult to obtain, we use only one or ten labelled anomaly from a randomly chosen anomaly type of each seen class during model training. In contrast to the previous work [7] which samples labelled anomalies from every anomaly type, our method achieves better data efficiency. The Area Under ROC Curve is used as the performance criteria and the reported experimental results are averaged over 10 trial runs.

**Text Prompts Formulation** We use the Oxford English Dictionary definition to construct the normality prompt for each object class. For example, for "*screw*" class, the normality prompt is "a short metal pin with a helical thread running round it and a slotted head". Oppositely, abnormality prompts describe our prior knowledge of the visual appearance of potential anomalies from that same class. Our abnormality prompt is constructed as: "abnormality of {*class label*} is [*anomaly description*]". Using our earlier example of "screw", example anomaly descriptions include "scratch, tear, crack, cut, defect", which are provided as anomaly type labels in the dataset [4]. Patches that align closely with such an abnormality prompt are likely to contain some type of anomaly, meaning the sample from which the patches come is likely to be anomalous.

**Implementation Details** We use the pre-trained CLIP (ViT-B/16) model for the image and text encoders, and the two branches consist of three transformer blocks. The parameters of two branches are initialized from scratch. The top-K setting is the same as in [7], which is set to be 10 percent of the number of last-layer output tokens. We do not use any image augmentation techniques nor pseudo anomaly samples. For model training, we learn our PromptAD in the support-query manner. To simulate the real-world ZSAD scenario, we sample a few query images with ground-truth labels in each training episode and utilize the two views of text prompts as support information. Training is performed on each transformer layer with the proposed symmetric deviation loss. To prevent the vast number of normal samples from overwhelming the training loss, we up-sample the positive samples (anomalies) by 10 times for balanced model training. After training, the model can be directly applied to novel classes without further updating.

### 4.2. Results

We compare PromptAD against the following baseline methods: DevNet [17], MLEP [14], DRA [7] , WinCLIP [10] and the original CLIP model [18].

**Classification** Table 1 shows the performance of PromptAD compared with other methods. Our method outperforms all its competitors on of all the unseen classes, and this is especially significant in the 1-shot setting, demonstrating that our method can effectively exploit rich semantic information in the two-branch learning framework to boost the anomaly detection performance. Furthermore, in more difficult object classes, such as the transistor images which contain many components and a complex background, the advantages of our proposed method become more significant. This may be attributed to the incorporation of text prompts, which provide important semantic information for generalization to unseen classes. In comparison, MLEP [14], DevNet [17] and DRA [7] rely on images solely and therefore do not generalise as successfully. The CLIP [18] baseline focuses more on semantic classification by aligning the relevant textual information with the image, failing to preserve the necessary spatial information to produce fine-grained and position-sensitive anomaly detection, which is resolved by the two-branch learning mechanism used in our model. In addition, we further

**Localization** We also show the pixel-wise AUC performance for anomaly localization in Table 2. We can see that PromptAD works well on anomaly localization and performs better than the state-of-the-art method WinCLIP [10]. This is because that PromptAD are explicitly designed for learning domain-specific features and thus ensures the vision-language correspondence on AD tasks. In contrast,

Table 1. Average AUC performance under the zero-shot anomaly detection setting for the MVTec dataset and the AITEX dataset. The best score is highlighted in red.

| **Dataset** | **One Anomaly Example** | | | | | **Ten Anomaly Examples** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **CLIP** [18] | **DevNet** [17] | **MLEP** [14] | **DRA** [7] | **Ours** | **CLIP** [18] | **DevNet** [17] | **MLEP** [14] | **DRA** [7] | **Ours** |
| **MVTec AD** | | | | | | | | | | |
| Carpet | 0.785 | 0.827 | 0.875 | 0.897 | **0.992** | 0.812 | 0.879 | 0.901 | 0.927 | **0.997** |
| Grid | 0.624 | 0.794 | 0.832 | 0.841 | **0.955** | 0.681 | 0.819 | 0.860 | 0.854 | **0.978** |
| Leather | 0.702 | 0.797 | 0.860 | 0.959 | **1.000** | 0.743 | 0.856 | 0.889 | 0.981 | **1.000** |
| Tile | 0.689 | 0.767 | 0.852 | 0.869 | **0.990** | 0.697 | 0.738 | 0.891 | 0.891 | **0.995** |
| Wood | 0.713 | 0.889 | 0.904 | 0.927 | **0.991** | 0.730 | 0.904 | 0.893 | 0.956 | **1.000** |
| Bottle | 0.602 | 0.807 | 0.913 | 0.864 | **0.979** | 0.661 | 0.824 | 0.865 | 0.897 | **0.992** |
| Capsule | 0.609 | 0.725 | 0.756 | 0.816 | **0.916** | 0.638 | 0.751 | 0.775 | 0.824 | **0.935** |
| Pill | 0.542 | 0.765 | 0.773 | 0.798 | **0.853** | 0.596 | 0.758 | 0.802 | 0.817 | **0.872** |
| Transistor | 0.565 | 0.586 | 0.691 | 0.727 | **0.812** | 0.599 | 0.612 | 0.751 | 0.743 | **0.831** |
| Zipper | 0.573 | 0.734 | 0.808 | 0.869 | **0.947** | 0.597 | 0.778 | 0.853 | 0.889 | **0.958** |
| Cable | 0.612 | 0.670 | 0.653 | 0.707 | **0.847** | 0.615 | 0.693 | 0.612 | 0.724 | **0.867** |
| Hazelnut | 0.839 | 0.906 | 0.931 | 0.949 | **0.985** | 0.861 | 0.938 | 0.950 | 0.973 | **1.000** |
| Metal nut | 0.521 | 0.605 | 0.645 | 0.689 | **0.824** | 0.583 | 0.627 | 0.665 | 0.716 | **0.845** |
| Screw | 0.551 | 0.523 | 0.576 | 0.619 | **0.698** | 0.567 | 0.619 | 0.609 | 0.642 | **0.712** |
| Toothbrush | 0.526 | 0.663 | 0.690 | 0.703 | **0.807** | 0.577 | 0.690 | 0.728 | 0.753 | **0.815** |
| **Average** | 0.630 | 0.737 | 0.784 | 0.815 | **0.908** | 0.663 | 0.765 | 0.802 | 0.839 | **0.912** |
| **AITEX** | 0.763 | 0.825 | 0.808 | 0.864 | **0.927** | 0.804 | 0.861 | 0.879 | 0.893 | **0.931** |

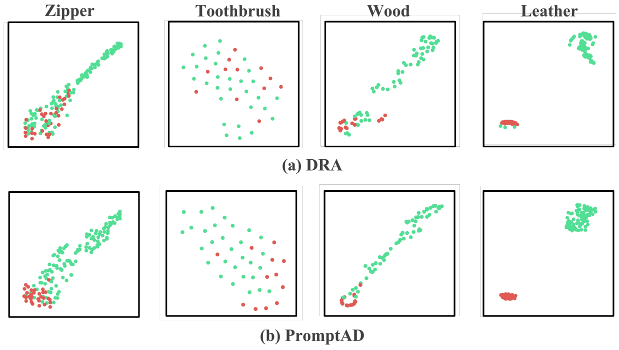
Table 2. Pixel-level AUC performance on MVTec AD dataset.

| Methods | pAUROC | PRO | $F_1$-max |
|---|---|---|---|
| **Trans-MM** [5] | 0.575 | 0.219 | 0.121 |
| **MaskCLIP** [25] | 0.637 | 0.405 | 0.185 |
| **DRA** [7] | 0.792 | 0.553 | 0.269 |
| **WinCLIP** [10] | 0.851 | 0.646 | 0.317 |
| **PromptAD** | **0.921** | **0.728** | **0.362** |

WinCLIP relies on features generated by the frozen CLIP model, which may not be optimal for anomaly localization. Some qualitative results are shown in Fig. 5.

Fig. 4 visualizes the representations learned by PromptAD in comparison to those by DRA [7]. We see that normal samples and anomalies exhibit greater separability in the latent space obtained through our dual-branch framework with semantic knowledge injected.

**Few-shot AD** We also evaluate our method under the general few-shot AD setting used by RegAD [9]. This setting assumes that a few normal examples of a new class are available for model testing. Therefore, we replace the normality prototype $\phi_{nor}$ with the visual embeddings of the available normal examples. The evaluation results are presented in Table 3. The sampled examples are then removed from the test set during evaluation. Under the case of K = 2, the performance gains observed in our proposed method over the other baselines of CLIP, RegAD [9] and DRA [7] are 18.0, 5.5, 12.0 points, respectively, demonstrating that our method can exploit the rich semantic information effectively to boost the overall detection performance. When the number of available visual samples increases, the gap between our method and the other methods become smaller. This may be because visual modal might be richer and more discriminative than text ones for AD task when more visual examples are available.



(a) DRA

(b) PromptAD

Figure 4. Visualizations of features learned by DRA and our proposed PromptAD on the unseen test classes of MVTec dataset. Green indicates anomalies while red indicates normal samples. Features learned by PromptAD show better separability.

## 4.3. Ablation Study

**Importance of Each Component** After validating the overall effectiveness of our approach, we further investigate the importance of each component. The results are presented in Table 4, where we denote the abnormality branch coupled with abnormality prompts as ABN and the normality branch coupled with normality prompts as NOR. Sim-

Table 3. Average AUC performance under the general anomaly detection setting on the MVTec AD dataset. "$K$-shot" denotes the number of normal samples used for training.

| $K$-shot | CLIP [18] | RegAD [9] | DRA [7] | PromptAD |
|---|---|---|---|---|
| 8 | 0.760 | 0.912 | 0.853 | **0.931** |
| 4 | 0.758 | 0.882 | 0.821 | **0.927** |
| 2 | 0.732 | 0.857 | 0.792 | **0.912** |

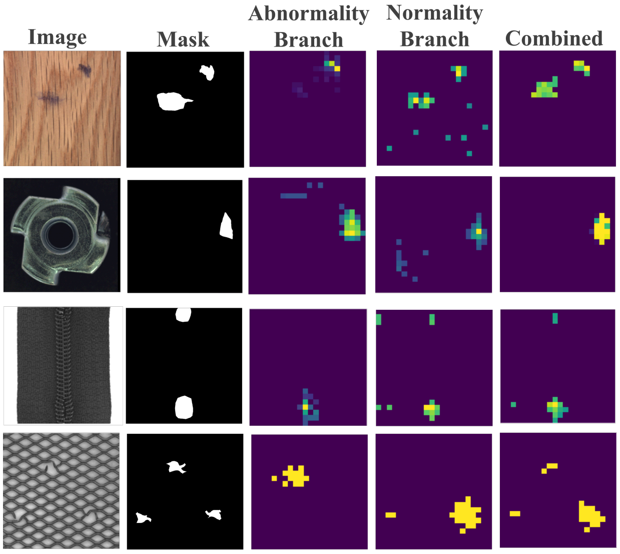


Figure 5. Qualitative anomaly localization results on MVTec AD dataset. For each example, the images from left to right are the anomaly image, the ground-truth mask, the anomaly score map produced by the abnormality branch and normality branch, and the fused score map.

Table 4. Ablation study for PromptAD. ABN and NOR are the abnormality and normality branches respectively.

| Components | 1-shot | 10-shot |
|---|---|---|
| ABN | 0.741 | 0.765 |
| NOR | 0.735 | 0.763 |
| ABN + NOR | 0.746 | 0.789 |
| ABN + NOR + CCL | 0.838 | 0.866 |
| ABN + NOR + CCL + CMI | **0.908** | **0.912** |

ply adopting the single-view training without any constraint leads to poor generalization performance due to semantic confusion. In contrast, the proposed CCL alleviates this issue, outperforming the naive single-view training approach up to 10.2 percentage points in the 10-shot scenario. We then compare the performances of the single-branch and double-branch frameworks. The results show that the double-branch framework outperforms the single branch by a significant margin since it takes into account both normal and anomaly data distribution. This is also validated by the visualizations in Sec. 4.3.

CMI contributes to forming a virtuous feedback cycle between the two branches via exchanging complementary information, and is thus able to improve the detection accuracy. As a result, one branch is able to learn better abstract representations by referring to the other branch's predictions. In contrast to the baselines where this feedback cycle is absent, either branch tends to overfit to its own view and suffers from significant performance degradation.

**Visualizing Detection Maps of Each Branch** To gain deeper insights into the anomaly detection capabilities of the abnormality branch (ABN Branch) and normality branch (NOR Branch) in PromptAD, we visualize the AD score maps from each branch in Fig. 5 for the unseen classes of "*wood*," "*metal nut*," "*zipper*," and "*grid*". We see the abnormality branch focuses on the most discriminative anomaly regions but does not identify all potential anomaly areas. In contrast, the normality Branch presents a more comprehensive segmentation mask, localizing most of the anomaly regions. This can be attributed to the guidance by normality semantic information, which allows it to identify visual patterns that deviate from the expected normality profile as anomalies. Moreover, we observe that score fusion effectively enhances the AD performance and mitigates the impact of incorrect detections (false positives). This validates that the discriminative capabilities of the abnormality and normality branches complement each other and both contribute to improved ZSAD performance through the dual-branch mechanism.

## 5. Conclusion

The closed-set nature of existing anomaly detection methods limit their generalization capabilities to new distributions, such as for previously unseen classes. To address this, we study the ZSAD problem where a model is trained to detect anomalies from seen classes and tested on its ability to detect anomalies in unseen classes without additional training. We propose PromptAD: a dual branch framework that incorporates rich semantic knowledge from both abnormality and normality views in the form of natural language text prompts. PromptAD refines the representations from a fixed CLIP encoder backbone for the anomaly detection task using the a dual-branch framework, each making use of abnormality and normality information in synergy. Importantly, information from one view is used to help and complement the learning of the other view, through cross-view contrastive learning and cross-view mutual interaction. Extensive experiments show that PromptAD improves zero-shot anomaly detection over existing baselines on key industrial benchmark datasets and can also maintain its strong performance in few-shot settings.

## References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018. 3

[2] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015. 3

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 5

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 7

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 8

[6] Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. Autoencoder-based network anomaly detection. In *2018 Wireless telecommunications symposium (WTS)*, pages 1–5. IEEE, 2018. 3

[7] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7388–7398, 2022. 3, 7, 8, 9

[8] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 3(7):e11, 2018. 4

[9] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael W. Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, volume 13684, pages 303–319. Springer, 2022. 7, 8, 9

[10] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 3, 7, 8

[11] Daiki Kimura, Subhajit Chaudhury, Minori Narita, Asim Munawar, and Ryuki Tachibana. Adversarial discriminative attention for robust anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2172–2181, 2020. 6

[12] Doyup Lee, Yeongjae Cheon, and Wook-Shin Han. Regularizing attention networks for anomaly detection in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1845–1853, 2021. 6

[13] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 3

[14] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, pages 3023–3030, 2019. 3, 7, 8

[15] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 4

[16] Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4578–4585, 2019. 6

[17] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 353–362, 2019. 3, 6, 7, 8

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 7, 8, 9

[19] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pages 4–11, 2014. 3

[20] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 3

[21] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*, pages 485–503. Springer, 2020. 6

[22] Runqi Wang, Hao Zheng, Xiaoyue Duan, Jianzhuang Liu, Yuning Lu, Tian Wang, Songcen Xu, and Baochang Zhang. Few-shot learning with visual distribution calibration and cross-modal distribution alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23445–23454, 2023. 2

[23] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[24] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018. 3

[25] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 8

[26] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017. 3

[27] Yuan Zhou, Yanrong Guo, Shijie Hao, Richang Hong, and Jiebo Luo. Few-shot partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6

[28] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8402–8411, 2021. 6