Toward Generalist Anomaly Detection via In-context Residual Learning with Few-shot Sample Prompts

Jiawen Zhu and Guansong Pang *

School of Computing and Information Systems, Singapore Management University

Abstract

This paper explores the problem of Generalist Anomaly Detection (GAD), aiming to train one single detection model that can generalize to detect anomalies in diverse datasets from different application domains without any further training on the target data. Some recent studies have showed that large pre-trained Visual-Language Models (VLMs) like CLIP have strong generalization capabilities on detecting industrial defects from various datasets, but their methods rely heavily on handcrafted text prompts about defects, making them difficult to generalize to anomalies in other applications, e.g.*, medical image anomalies or semantic anomalies in natural images. In this work, we propose to train a GAD model with few-shot normal images as sample prompts for AD on diverse datasets on the fly. To this end, we introduce a novel approach that learns an in-context residual learning model for GAD, termed* `InCTRL`*. It is trained on an auxiliary dataset to discriminate anomalies from normal samples based on a holistic evaluation of the residuals between query images and few-shot normal sample prompts. Regardless of the datasets, per definition of anomaly, larger residuals are expected for anomalies than normal samples, thereby enabling* `InCTRL` *to generalize across different domains without further training. Comprehensive experiments on nine AD datasets are performed to establish a GAD benchmark that encapsulate the detection of industrial defect anomalies, medical anomalies, and semantic anomalies in both one-vs-all and multi-class setting, on which* `InCTRL` *is the best performer and significantly outperforms state-of-the-art competing methods. Code is available at https://github.com/mala-lab/InCTRL.*

1. Introduction

Anomaly Detection (AD) is a crucial computer vision task that aims to detect samples that substantially deviate from the majority of samples in a dataset, due to its broad real-life

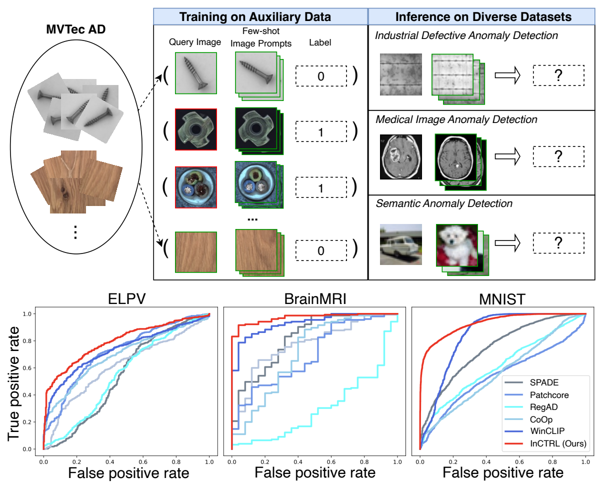

Figure 1. **Top**: An illustration of `InCTRL`: a one-for-all model using few-shot normal images as sample prompts. **Bottom**: AUROC curves of `InCTRL` and competing few-shot methods on three different application datasets without any training on the target data.

applications such as industrial inspection, medical imaging analysis, and scientific discovery, etc. [12, 36]. Current AD paradigms are focused on individually building one model on the training data, *e.g.*, a set of anomaly-free samples, of each target dataset, such as data reconstruction approach [2, 22, 31, 38, 43, 60, 63–65, 67–69], one-class classification [6, 15, 41, 48, 66], and knowledge distillation approach [8, 11, 19, 42, 51, 52, 70]. Although these approaches have shown remarkable detection performance on various AD benchmarks, they require the availability of large training data and the skilled detection model training per dataset. Thus, they become infeasible in application scenarios where training on the target dataset is not allowed due to either data privacy issues, *e.g.*, arising from using those data in training the models due to machine unlearning [62], or unavailability of large-scale training data in the deployment of new applications. To tackle these challenges, this paper explores the problem of learning **Generalist Anomaly Detection** (GAD) models, *aiming to train*

*Corresponding author: G. Pang (gspang@smu.edu.sg)

one single detection model that can generalize to detect anomalies in diverse datasets from different application domains without any training on the target data.

Being pre-trained on web-scale image-text data, large Visual-Language Models (VLMs) like CLIP [39] have exhibited superior generalization capabilities in recent years, achieving accurate visual recognition across different datasets without any fine-tuning or adaptation on the target data. More importantly, some very recent studies (*e.g*., WinCLIP [25]) show that these VLMs can also be utilized to achieve remarkable generalization on different defect detection datasets. Nevertheless, a significant limitation of these models is their dependency on a large set of manually crafted prompts specific to defects. This reliance restricts their applicability, making it challenging to extend their use to detecting anomalies in other data domains, *e.g*., medical image anomalies [10, 20, 42, 49, 50] or semantic anomalies in one-vs-all or multi-class settings [11, 41].

To address this problem, we propose to train a GAD model that aims to utilize few-shot normal images from any target dataset as sample prompts for supporting GAD on the fly, as illustrated in Figure 1(Top). The few-shot setting is motivated by the fact that it is often easy to obtain few-shot normal images in real-world applications. Furthermore, these few-shot samples are not used for model training/tuning; they are just used as sample prompts for enabling the anomaly scoring of test images during inference. This formulation is fundamentally different from current few-shot AD methods [5, 23, 44, 45, 56, 57, 61] that use these target samples and their extensive augmented versions to train the detection model, which can lead to an overfitting of the target dataset and fail to generalize to other datasets, as shown in Figure 1(Bottom).

We then introduce an GAD approach, the first of its kind, that learns an in-context residual learning model based on CLIP, termed `InCTRL`. It trains an GAD model to discriminate anomalies from normal samples by learning to identify the residuals/discrepancies between query images and a set of few-shot normal images from auxiliary data. The few-shot normal images, namely *in-context sample prompts*, serve as prototypes of normal patterns. When comparing with the features of these normal patterns, per definition of anomaly, a larger residual is typically expected for anomalies than normal samples in datasets of different domains, so the learned in-context residual model can generalize to detect diverse types of anomalies across the domains. To capture the residuals better, `InCTRL` models the in-context residuals at both the image and patch levels, gaining an in-depth in-context understanding of what constitutes an anomaly. Further, our in-context residual learning can also enable a seamless incorporation of normal/abnormal text prompt-guided prior knowledge into the detection model, providing an additional strength for the detection from the text-image-aligned semantic space.

Accordingly, we make the following main contributions.

- We introduce a GAD task to evaluate the generalization capability of AD methods in identifying anomalies across various scenarios without needing to training/tuning on the target datasets. To the best of our knowledge, this is the first study dedicated to a generalist approach to anomaly detection, encompassing industrial defects, medical anomalies, and semantic anomalies.
- We then propose an in-context residual learning framework for GAD, called `InCTRL`. It is designed to distinguish anomalies from normal samples by detecting residuals between test images and in-context few-shot normal sample prompts from any target dataset on the fly. `InCTRL` is optimized on auxiliary data to achieve the one-model-for-all goal, *i.e*., one model for AD on diverse datasets without any training on target data.
- Extensive experiments on nine AD datasets are performed to establish a GAD benchmark that encapsulates three types of popular AD tasks, including industrial defect anomaly detection, medical image anomaly detection, and semantic anomaly detection under both one-vs-all and multi-class settings. Our results show that `InCTRL` significantly surpasses existing state-of-the-art methods.

2. Related Work

2.1. Anomaly Detection

Anomaly Detection (AD). Existing AD approaches typically rely on unsupervised learning due to the scarcity of anomaly data. Numerous methods have been introduced. One-class classification methods [6, 15, 41, 48, 66] focus on compactly describing normal data with support vectors. Reconstruction-based methods [2, 22, 31, 38, 43, 60, 63–65, 67–69] train models to reconstruct normal images, where anomalies are identified by higher reconstruction errors. Distance-based methods [16, 17, 40] determine anomalies based on the distance between test image embeddings and normal reference embeddings from stored training data. Knowledge distillation methods [8, 11, 19, 42, 51, 52, 70] focus on distilling normal patterns from pre-trained models and detect anomalies based on the difference between distilled and original features. The above approaches are designed to fit on target dataset for AD, *i.e*., one model for one dataset, while we aim for a one-model-for-all setting. A relevant research line is to tackle the AD problem under domain or distribution shift [1, 11, 20, 33, 65, 73], but they generally assume a large domain relevance on the source and target data. Additionally, there have been a number of concurrent studies leveraging VLMs for AD [58, 59, 72], but they address a different setting from ours, *e.g*., weakly-supervised AD [58, 59] or zero-shot AD [72].

Few-shot Anomaly Detection (FSAD). FSAD is designed to identify anomalies using limited normal samples from target datasets. Traditional FSAD research focuses on modeling the normal distribution of these few normal samples to detect anomalies [5, 23, 29, 44, 45, 56, 57, 61]. However, these methods often cannot generalize to new domains, as they generally require re-training or fine-tuning with normal data from target datasets. Distance-based approaches such as SPADE [16], PaDiM [17] and PatchCore [40] present a solution to address this problem by making full use of pre-trained representations of the few-shot samples to calculate distance-based anomaly scores without training. Recently, RegAD [23] is designed to operate without the need for training or fine-tuning on new data for the FSAD task, but it requires domain relevance between training and test data to work well. WinCLIP [25] pioneers the application of large Visual-Language Models (VLM) on zero-shot and few-shot AD tasks by processing images through multi-scale window movements and text prompting to CLIP. Without adapting CLIP to the AD task, WinCLIP gains impressive detection performance on defect datasets using its handcrafted text prompts, but it fails to work well when the text prompts cannot capture the required anomaly semantics, making it difficult to generalize well to diverse AD tasks.

2.2. In-Context Learning

In-Context Learning (ICL) is an innovative approach that helps enhance the performance of Large Language Models (LLMs) in Natural Language Processing (NLP) [3, 9, 21], which leverages minimal in-context prompts to adapt LLMs to novel tasks effectively. Recently, several studies [13, 14, 26, 32, 53] attempt to apply ICL to vision tasks by converting vision problems to NLP ones using the language or specially-designed discrete tokens as the task prompts. conversely, Amir *et al.* [4] introduce a novel approach for in-context visual prompting by treating a spectrum of vision tasks as grid in-painting problems. Similarly, Painter [54, 55] then proposes to perform masked image in-painting. However, these methods focus more on task-level generalization, so they are not applicable to the AD task which focuses more on the instance-level discrepancy. Our work redesign ICL for GAD. We redefine few-shot image prompts as dataset-specific normal patterns, rather than as an instruction for particular tasks. By capturing the in-context residual between the query images and their normal patterns, our model can gain a cohesive understanding of diverse anomalies, enabling remarkable generalized detection performance for GAD.

3. `InCTRL`: In-Context Residual Learning

3.1. Problem Statement

The objective of GAD is to train a single AD model that works well for detecting anomalies on test datasets from diverse application domains without any training on the target data. Thus, the training set is assumed to be drawn from different distributions from the test sets. Formally, let $\mathcal{D}_{train} = \{X_{train}, Y_{train}\}$ be an *auxiliary* training dataset with normal and anomaly class labels, where $X_{train} = \{x_i\}_{i=1}^N$ consists of N normal and anomalous images and $Y_{train} = \{y_i\}_{i=1}^N$, with $y_i = 0$ indicates normal and $y_i = 1$ signifies abnormal. A collection of test sets, $\mathcal{T} = \{\mathcal{D}_{test}^1, \mathcal{D}_{test}^2, \cdots, \mathcal{D}_{test}^M\}$ with $\mathcal{D}_{test}^j = \{X_{test}^j, Y_{test}^j\}$, from M different application domains with various types of anomalies is given. The test sets are drawn from a distribution different from that of $\mathcal{D}_{train}$. Then the goal is to train a generalist anomaly scoring function: $\mathcal{D}_{train} \to \mathbb{R}$ so that it assigns larger anomaly scores to the anomalous samples than to the normal ones from any test dataset in $\mathcal{T}$. In the context of GAD with few-shot normal samples, a small set of a few normal images randomly drawn from the target domain, $\mathcal{P} = \{p_1, p_2, \cdots, p_K\}$ where K is typically a small number, *e.g.*, $K \ll N$, is available during inference, but $\mathcal{P}$ is not available in any way during the training of the generalist detection model.

3.2. Overview of Our Approach `InCTRL`

Our approach `InCTRL` is designed to effectively model the in-context residual between a query image and a set of few-shot normal images as sample prompts, utilizing the generalization capabilities of CLIP to detect unusual residuals for anomalies from different application domains. CLIP is a VLM consisting of a text encoder $f_t(\cdot)$ and a visual encoder $f_v(\cdot)$, with the image and text representations from these encoders well aligned by pre-training on web-scale text-image data. `InCTRL` is optimized using auxiliary data $\mathcal{D}_{train}$ via an in-context residual learning in the image encoder, with the learning augmented by text prompt-guided prior knowledge from the text encoder.

To be more specific, as illustrated in Fig. 2, we first simulate an in-context learning example that contains one query image x and a set of few-shot normal sample prompts $\mathcal{P}'$, both of which are randomly sampled from the auxiliary data $\mathcal{D}_{train}$. Through the visual encoder, we then perform multi-layer patch-level and image-level residual learning to respectively capture local and global discrepancies between the query and few-shot normal sample prompts (Secs. 3.3 and 3.4). Further, our model allows a seamless incorporation of normal and abnormal text prompts-guided prior knowledge from the text encoder based on the similarity between these textual prompt embeddings and the query images (Sec. 3.5). The training of `InCTRL` is to optimize a few projection/adaptation layers attached to the visual encoder to learn a larger anomaly score for anomaly samples than normal samples in $\mathcal{D}_{train}$, with the original parameters in both encoders frozen; during inference, a test image, together with the few-shot normal image prompts from the

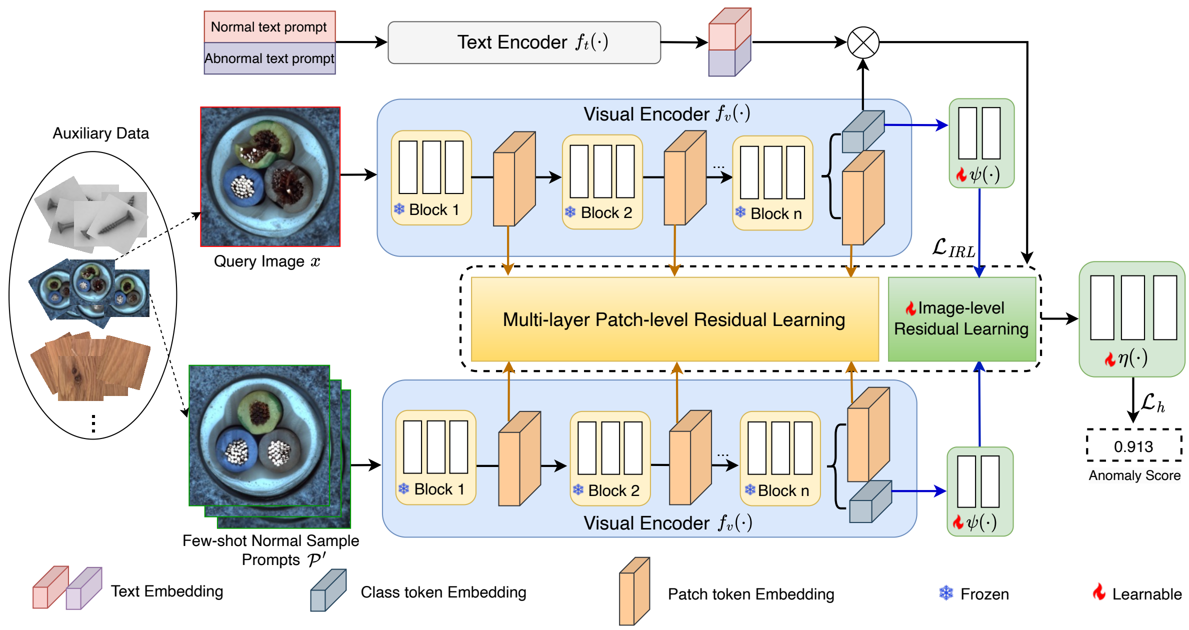

Figure 2. Overview of the training of InCTRL. It firstly simulates in-context learning scenarios using a query image and a few-shot normal sample prompts randomly drawn from the auxiliary training data. Then it performs multi-layer patch-level and image-level residual learning to capture both local and global residuals between the query image and the normal prompts. Lastly, those residual information, combined with text prompts-guided prior knowledge from the text encoder, is utilized for a holistic anomaly score learning.

target dataset and the text prompts, is put forward through our adapted CLIP-based GAD network, whose output is the anomaly score for the test image (Sec. 3.6). Below we present these modules in detail.

3.3. Multi-Layer Patch-Level Residual Learning

To effectively capture fine-grained in-context residuals between the query image and the normal image prompts, we introduce a multi-layer patch-level residual learning component in InCTRL. Typically, the CLIP visual encoder comprises a series of block layers. From the bottom to the top of layers, the visual encoder gradually learns the visual patterns at different levels of abstraction [39]. Thus, this component is designed to model patch-level in-context residuals from the patch token embeddings obtained from the multiple levels of the blocks within the visual encoder.

To be specific, assuming the visual encoder consists of n blocks, for a given set of few-shot normal sample prompts $\mathcal{P}'$ and a training query image x, we extract a series of patch token embedding maps $\{T_x^l\}_{l=1}^n$ and $\{T_{x'}^l\}_{l=1}^n$ where $T_{(\cdot)}^l \in \mathbb{R}^{h\times w\times d}$ and $x' \in \mathcal{P}'$, with h, w, and d be the height, width, and dimension of the feature map T respectively. At each layer l, the patch-level in-context residuals are captured by distances between the embeddings of the query token and the image prompt token across all image prompts in $\mathcal{P}'$. Formally, for the query image x, its multi-layer patch-level in-context residuals at layer l are modeled by a residual map $\mathbf{M}_x^l \in \mathbb{R}^{h\times w}$, where the residual value of each patch of x is calculated based on its patch embedding and the nearest patch embedding of all images in $\mathcal{P}'$ as:

$$\mathbf{M}_x^l(i,j) = 1 - \langle T_x^l(i,j), h(T_x^l(i,j)|\mathcal{P}')\rangle, \quad (1)$$

where $h(T_x^l(i,j)|\mathcal{P}')$ returns the embedding of the patch token that is most similar to $T_x^l(i,j)$ among all image patches in $\mathcal{P}'$, and $\langle\cdot\rangle$ is the cosine similarity function. The final patch-level residual map $\mathbf{M}_x \in \mathbb{R}^{h\times w}$ is averaged over n layer-wise residual maps:

$$\mathbf{M}_x = \frac{1}{n}\sum_{l=1}^{n}\mathbf{M}_x^l. \quad (2)$$

Each residual value in $\mathbf{M}_x$ is similar to a nearest-neighbor-distance anomaly score of the query patch to the image patch set in $\mathcal{P}'$. As shown in prior studies [16, 17, 34, 35, 40], such distance-based anomaly scores can effectively discriminate anomalies from normal samples. Thus, the resulting residual map $\mathbf{M}_x$ provides a feature set of collective anomaly-discriminative power at multi-layer resolutions for the subsequent anomaly score learning in InCTRL.

3.4. Image-level Residual Learning

In addition to the discriminative power at the local patch-level residuals, the global discriminative information at the image level is also significant and serve as complementary knowledge to the patch-level features.

Hence, we introduce an image-level residual learning component to capture the higher-level discrepancies between x and $\mathcal{P}'$. Intuitively, the class token embedding from the last block of the visual encoder is used as the feature input, as it captures the most image-level discriminative information due to the bottom-up abstraction of information in the visual encoder. However, it is important to note that CLIP was originally designed for classification tasks, focusing on the semantic of the objects in the scenery, which does not align well with the anomaly detection task in which both normal and abnormal samples are often from the same class of object. To reconcile this, we include an adapter layer $\psi(\cdot)$, parameterized by Θ_ψ, to adapt the image representations further to anomaly detection, and thus, we learn the image-level residuals based on the adapted image features. Further, the prototypical features of the few-shot sample prompts, rather than the features of individual sample, are used to learn the in-context residuals, since they help capture more representative features of normal patterns.

Specifically, let $f_v(x) \in \mathbb{R}^{d'}$ be the class token embedding of input x in the visual encoder, we first compute the prototype of the feature maps of the image prompts in $\mathcal{P}'$:

$$\mathbf{I}_p = \frac{1}{K} \sum_{x'_k \in \mathcal{P}'} \psi(f_v(x'_k); \Theta_\psi), \tag{3}$$

where $\mathbf{I}_p \in \mathbb{R}^{d'}$. Then let $\mathbf{I}_x = \psi(f_v(x); \Theta_\psi)$ be the adapted features of the query image x, the in-context image-level residual features $\mathbf{F}_x$ for x are obtained by performing element-wise subtraction between two feature maps:

$$\mathbf{F}_x = \mathbf{I}_x \ominus \mathbf{I}_p, \tag{4}$$

where $\ominus$ denotes element-wise subtraction. Subsequently, these in-context residual features are fed to an image-level anomaly classification learner $\eta : \mathbf{F}_x \to \mathbb{R}$, parameterized by Θ_η which is optimized by the binary classification loss:

$$\mathcal{L}_{IRL} = \frac{1}{N} \sum_{x \in X_{train}} \mathcal{L}_b(\eta(\mathbf{F}_x; \Theta_\eta), y_x), \tag{5}$$

where $\mathcal{L}_b$ is a binary classification loss. Focal loss [30] is used by default in our model.

3.5. Fusing Text Prompt-based Prior Knowledge

The above two components are focused on residual learning based on the visual encoder. `InCTRL` also allows easy incorporation of text-prompt-guided prior knowledge about normality and abnormality from the text encoder of CLIP. This helps `InCTRL` leverage the normal and abnormal semantics hidden in the CLIP's pre-trained image-text-aligned embedding space for GAD. Motivated by this, `InCTRL` exploits the text encoder to extract text prompt-guided discriminative features. Since the text prompts designed in WinCLIP [25] show remarkable detection performance, `InCTRL` adopts the same text prompt templates and its ensemble strategy, including both state and template-level text prompts. At the state level, generic text descriptions are employed to differentiate between normal and abnormal objects, whereas the template level provides a list of specific prompts tailored for anomaly detection (see `Appendix B.3` for detailed text prompts used).

It should be noted that, unlike WinCLIP that uses these text prompts to directly compute the anomaly score, `InCTRL` utilizes them to extract text-prompt-guided features for complementing the patch- and image-level residual features obtained through the visual encoder.

Specifically, let $\mathcal{P}_t^n$ be the set of text prompts for the normal class, we use the prototype of the text prompt embeddings to provide a representative embedding of the normal text prompts $\mathbf{F}_n = \frac{1}{|\mathcal{P}_t^n|} \sum_{p_i \in \mathcal{P}_t^n} f_t(p_i)$ where $p_i \in \mathcal{R}^{d'}$; similarly we can obtain the prototype embedding for the abnormality text prompt set $\mathcal{P}_t^a$ by $\mathbf{F}_a = \frac{1}{|\mathcal{P}_t^a|} \sum_{p_j \in \mathcal{P}_t^a} f_t(p_j)$. Then, `InCTRL` extracts an AD-oriented discriminative feature based on the similarity between the query image x and the two prototypes of the text prompts:

$$s_a(x) = \frac{\exp(\mathbf{F}_a^\intercal f_v(x))}{\exp(\mathbf{F}_n^\intercal f_v(x)) + \exp(\mathbf{F}_a^\intercal f_v(x))}, \tag{6}$$

where $[\cdot]^\intercal$ denotes a transpose operation, and $s_a(x)$ is the probability of the input x being classified as abnormal.

3.6. Training and Inference

In-Context Residual Learning. During training, `InCTRL` performs a holistic residual learning that synthesizes both patch-level and image-level residual information, augmented by the text prompt-guided features. The holistic in-context residual map of a query image x is defined as:

$$\mathbf{M}_x^+ = \mathbf{M}_x \oplus s_i(x) \oplus s_a(x), \tag{7}$$

where $s_i(x) = \eta(\mathbf{F}_x; \Theta_\eta)$ is an anomaly score based on the image-level residual map $\mathbf{F}_x$ and $\oplus$ denotes an element-wise addition. `InCTRL` then devises a holistic anomaly scoring function ϕ, parameterized by Θ_ϕ, based on $\mathbf{M}_x^+$, and defines the final anomaly score as:

$$s(x) = \phi(\mathbf{M}_x^+; \Theta_\phi) + \alpha s_p(x), \tag{8}$$

where $\phi(\mathbf{M}_x^+; \Theta_\phi)$ performs a holistic anomaly scoring using patch-, image-level and text prompt-guided features, while $s_p(x) = \max(\mathbf{M}_x)$ is a maximum residual score-based fine-grained anomaly score at the image patch level. $s_p(x)$ is added into Eq. 8 because such patch-level anomaly scores are crucial for detecting local abnormal regions to which the ϕ-based holistic anomaly score can often overlook. α is a hyper-parameter that modulates the contribution of the patch-level residual score. Lastly, we optimize

Setup	Methods	Industrial Defects					Medical Anomalies		Semantic Anomalies			
									One-vs-all		Multi-class	
		ELPV	SDD	AITEX	VisA	MVTec AD	BrainMRI	HeadCT	MNIST	CIFAR-10	MNIST	CIFAR-10
Baseline (0-shot)		0.733±0.000	0.946±0.000	0.733±0.000	0.781±0.000	0.912±0.000	0.926±0.000	0.900±0.000	0.678±0.000	0.924±0.000	0.620±0.000	0.900±0.000
2-shot	SPADE	0.517±0.012	0.729±0.041	0.727±0.004	0.795±0.045	0.817±0.054	0.754±0.048	0.645±0.034	0.779±0.024	0.823±0.014	0.595±0.060	0.655±0.042
	PaDiM	0.594±0.083	0.721±0.015	**0.784±0.028**	0.680±0.042	0.785±0.025	0.657±0.122	0.595±0.036	-	-	-	-
	Patchcore	0.716±0.031	0.902±0.006	0.739±0.017	0.817±0.028	0.858±0.034	0.706±0.009	0.736±0.096	0.756±0.004	0.602±0.009	0.603±0.009	0.703±0.008
	RegAD	0.571±0.016	0.499±0.008	0.564±0.072	0.557±0.053	0.640±0.047	0.449±0.129	0.602±0.018	0.525±0.030	0.534±0.005	0.608±0.026	0.695±0.002
	CoOp	0.762±0.011	0.897±0.006	0.687±0.062	0.806±0.023	0.858±0.016	0.725±0.020	0.811±0.003	0.557±0.006	0.527±0.011	0.612±0.007	0.393±0.009
	WinCLIP	**0.726±0.020**	**0.942±0.006**	0.726±0.055	**0.842±0.024**	**0.931±0.019**	**0.934±0.012**	**0.915±0.015**	**0.810±0.008**	**0.925±0.001**	**0.632±0.000**	**0.914±0.005**
	Ours (`InCTRL`)	**0.839±0.003**	**0.972±0.011**	**0.761±0.029**	**0.858±0.022**	**0.940±0.015**	**0.973±0.027**	**0.929±0.025**	**0.892±0.009**	**0.935±0.002**	**0.635±0.010**	**0.924±0.005**
4-shot	SPADE	0.537±0.013	0.731±0.020	0.718±0.011	0.811±0.040	0.828±0.044	0.759±0.070	0.624±0.012	0.810±0.009	0.836±0.006	0.588±0.041	0.631±0.063
	PaDiM	0.612±0.080	0.742±0.014	**0.787±0.038**	0.735±0.031	0.805±0.018	0.792±0.048	0.622±0.013	-	-	-	-
	Patchcore	0.756±0.073	0.923±0.008	0.733±0.002	0.843±0.025	0.885±0.026	0.794±0.040	0.805±0.006	0.833±0.009	0.639±0.010	0.497±0.044	0.739±0.011
	RegAD	0.596±0.040	0.525±0.027	0.596±0.074	0.574±0.042	0.663±0.032	0.571±0.149	0.522±0.050	0.548±0.053	0.534±0.002	0.596±0.075	0.677±0.161
	CoOp	**0.781±0.002**	0.902±0.006	0.720±0.017	0.818±0.018	0.874±0.017	0.759±0.033	0.860±0.032	0.563±0.004	0.537±0.005	0.618±0.002	0.395±0.008
	WinCLIP	0.754±0.009	**0.943±0.004**	0.764±0.025	**0.858±0.025**	**0.940±0.021**	**0.941±0.002**	**0.912±0.003**	**0.851±0.010**	**0.927±0.001**	**0.632±0.004**	**0.915±0.003**
	Ours (`InCTRL`)	**0.846±0.011**	**0.975±0.006**	**0.790±0.018**	**0.877±0.019**	**0.945±0.018**	**0.975±0.016**	**0.933±0.013**	**0.902±0.016**	**0.940±0.010**	**0.643±0.007**	**0.928±0.009**
8-shot	SPADE	0.567±0.034	0.741±0.011	0.708±0.006	0.821±0.042	0.840±0.057	0.794±0.039	0.626±0.022	0.829±0.009	0.849±0.006	0.597±0.028	0.656±0.037
	PaDiM	0.724±0.017	0.769±0.037	0.792±0.025	0.768±0.032	0.820±0.016	0.758±0.025	0.661±0.039	-	-	-	-
	Patchcore	**0.837±0.016**	0.925±0.003	0.745±0.002	0.860±0.026	0.922±0.019	0.812±0.016	0.817±0.034	**0.876±0.004**	0.672±0.006	0.526±0.019	0.764±0.004
	RegAD	0.633±0.027	0.594±0.029	0.603±0.062	0.589±0.040	0.674±0.033	0.632±0.079	0.628±0.026	0.547±0.063	0.555±0.008	0.573±0.076	0.587±0.211
	CoOp	0.817±0.012	0.898±0.005	0.769±0.008	0.822±0.021	0.880±0.014	0.755±0.003	0.914±0.027	0.567±0.007	0.542±0.005	0.619±0.004	0.399±0.006
	WinCLIP	0.814±0.010	**0.941±0.001**	**0.796±0.015**	**0.868±0.020**	**0.947±0.025**	**0.944±0.001**	**0.915±0.008**	0.867±0.007	**0.928±0.001**	**0.641±0.004**	**0.916±0.003**
	Ours (`InCTRL`)	**0.872±0.013**	**0.978±0.006**	**0.806±0.036**	**0.887±0.021**	**0.953±0.013**	**0.983±0.012**	**0.936±0.008**	**0.920±0.003**	**0.945±0.002**	**0.646±0.003**	**0.934±0.008**

Table 1. AUROC results(mean±std) on nine real-world AD datasets under various few-shot AD settings. Best results and the second-best results are respectively highlighted in **red** and **blue**. 'Baseline' is a WinCLIP-based zero-shot AD model.

the final anomaly score $s(x)$ using X_{train}:

$$\mathcal{L}_h = \frac{1}{N} \sum_{x \in X_{train}} \mathcal{L}_b(s(x), y_x). \tag{9}$$

Thus, the full `InCTRL` model is optimized by minimizing the overall loss as follows:

$$\mathcal{L}_{InCTRL} = \mathcal{L}_{IRL} + \mathcal{L}_h. \tag{10}$$

Inference. During inference, for a given test image x_t and the K-shot normal image prompt set $\mathcal{P}$ from the target dataset, they are fed forward through the visual encoder and the adapter layers, obtaining $\mathbf{M}_{x_t}$ and $s_i(x_t)$. The text prompt sets used during training are used to obtain $s_a(x_t)$. Lastly, we obtain the final anomaly score of x_t via Eq. 8.

4. Experiments

4.1. Experimental Setup

Datasets. To verify the efficiency of our method `InCTRL`, we conduct comprehensive experiments across nine real-world AD datasets, including five industrial defect inspection dataset (MVTec AD [7], VisA [74], AITEX [46], ELPV [18], SDD [47]), two medical image datasets (BrainMRI [42], HeadCT [42]), and two semantic anomaly detection datasets: MNIST [28] and CIFAR-10 [27] under both one-vs-all and multi-class protocols [11, 41]. Under the one-vs-all protocol, one class is used as normal, with the other classes treated as abnormal; while under the multi-class protocol, images of even-number classes from MNIST and animal-related classes from CIFAR-10 are treated as normal, with the images of the other classes are considered as anomalies (see `Appendix A` for more details).

To assess the GAD performance, MVTec AD, the combination of its training and test sets, is used as the auxiliary training data, on which GAD models are trained, and they are subsequently evaluated on the test set of the other eight datasets without any further training. We train the model on VisA when evaluating the performance on MVTec AD. The few-shot normal prompts for the target data are randomly sampled from the training set of target datasets and remain the same for all models for fair comparison. We evaluate the performance with the number of few-shot normal prompt set to $K = 2, 4, 8$. The reported results are averaged over three independent runs with different random seeds.

Competing Methods and Evaluation Metrics. Since we aim to achieve a generalist AD model, the comparision is focus on detectors of similar generalist detection capabilities. Following [25], `InCTRL` is compared with three conventional full-shot AD approaches, including SPADE [16], PaDiM [17], and PatchCore [40], all of which are adapted to the few-shot setting by performing their distance-based anomaly scoring based on the few-shot normal samples. We also compare with state-of-the-art (SotA) conventional few-shot AD method RegAD [23] and the CLIP-driven method WinCLIP [25]. The popular prompt learning method CoOp [71] is used as an additional baseline that is trained on the auxiliary data as `InCTRL`, after which it uses the few-shot anomaly scoring strategy in WinCLIP to perform anomaly detection.

As for evaluation metrics, following previous works [11, 23, 25, 37, 40], we use two popular metrics AUROC (Area Under the Receiver Operating Characteristic) and AUPRC (Area Under the Precision-Recall Curve) to evaluate the AD performance. We also evaluate the number of parameters and per-image inference time of CLIP-based methods, which is presented in our `Appendix C.1`.

Implementation Details. By default, for CLIP-based models, including WinCLIP, CoOp and our `InCTRL`, we adopt the same CLIP implementation, OpenCLIP [24], and its

Setup	Methods	Industrial Defects					Medical Anomalies		Semantic Anomalies			
									One-vs-all		Multi-class	
		ELPV	SDD	AITEX	VisA	MVTec AD	BrainMRI	HeadCT	MNIST	CIFAR-10	MNIST	CIFAR-10
Baseline (0-shot)		0.855±0.000	0.886±0.000	0.552±0.000	0.812±0.000	0.957±0.000	0.988±0.000	0.970±0.000	0.940±0.000	0.990±0.000	0.606±0.000	0.852±0.000
2-shot	**SPADE**	0.618±0.007	0.366±0.105	0.470±0.008	0.818±0.031	0.922±0.023	0.952±0.009	0.851±0.022	**0.965±0.004**	0.971±0.003	**0.615±0.068**	0.502±0.035
	PaDiM	0.707±0.058	0.337±0.008	**0.529±0.034**	0.719±0.027	0.890±0.015	0.902±0.046	0.876±0.017	-	-	-	-
	Patchcore	0.840±0.031	0.676±0.003	0.378±0.008	0.841±0.023	0.939±0.012	0.921±0.017	0.913±0.002	0.956±0.001	0.926±0.002	0.482±0.025	0.574±0.015
	RegAD	0.679±0.005	0.173±0.019	0.275±0.035	0.614±0.037	0.837±0.034	0.872±0.065	0.854±0.009	0.913±0.006	0.909±0.003	0.612±0.013	0.672±0.008
	CoOp	0.841±0.020	0.543±0.004	0.443±0.050	0.835±0.019	0.922±0.007	0.923±0.002	0.937±0.014	0.926±0.003	0.911±0.002	0.607±0.009	0.371±0.013
	WinCLIP	**0.849±0.010**	**0.865±0.004**	0.500±0.043	**0.859±0.021**	**0.965±0.007**	**0.989±0.003**	**0.975±0.012**	0.963±0.001	**0.990±0.001**	0.614±0.005	**0.876±0.016**
	Ours (`InCTRL`)	**0.913±0.008**	**0.917±0.009**	**0.519±0.022**	**0.877±0.016**	**0.969±0.004**	**0.994±0.013**	**0.981±0.013**	**0.975±0.004**	**0.992±0.000**	**0.618±0.012**	**0.899±0.010**
4-shot	**SPADE**	0.627±0.011	0.385±0.018	0.451±0.031	0.826±0.024	0.924±0.015	0.958±0.017	0.854±0.016	0.966±0.008	0.973±0.002	**0.611±0.053**	0.487±0.047
	PaDiM	0.724±0.067	0.351±0.012	**0.540±0.053**	0.758±0.018	0.909±0.013	0.956±0.011	0.890±0.011	-	-	-	-
	Patchcore	**0.871±0.042**	0.703±0.013	0.377±0.001	0.860±0.016	0.950±0.013	0.945±0.017	0.941±0.009	**0.972±0.002**	0.934±0.003	0.504±0.025	0.606±0.010
	RegAD	0.688±0.018	0.176±0.003	0.294±0.031	0.628±0.034	0.846±0.026	0.900±0.041	0.810±0.028	0.916±0.013	0.908±0.001	0.522±0.085	0.681±0.127
	CoOp	0.867±0.003	0.594±0.014	0.454±0.014	0.842±0.016	0.924±0.008	0.932±0.013	0.957±0.017	0.929±0.002	0.915±0.003	0.611±0.003	0.374±0.012
	WinCLIP	0.864±0.004	**0.868±0.003**	0.513±0.017	**0.875±0.023**	**0.968±0.008**	**0.990±0.001**	**0.974±0.002**	0.971±0.002	**0.990±0.000**	**0.611±0.011**	**0.882±0.009**
	Ours (`InCTRL`)	**0.916±0.009**	**0.924±0.015**	**0.548±0.016**	**0.902±0.027**	**0.972±0.006**	**0.994±0.013**	**0.984±0.011**	**0.980±0.007**	**0.992±0.004**	**0.620±0.004**	**0.901±0.020**
8-shot	**SPADE**	0.641±0.018	0.394±0.024	0.427±0.008	0.844±0.031	0.930±0.016	0.962±0.014	0.860±0.019	0.974±0.002	0.976±0.001	0.613±0.035	0.515±0.024
	PaDiM	0.798±0.014	0.384±0.045	0.555±0.031	0.781±0.024	0.927±0.012	0.946±0.007	0.896±0.009	-	-	-	-
	Patchcore	**0.915±0.007**	0.708±0.009	0.389±0.003	0.873±0.022	0.962±0.013	0.957±0.007	0.931±0.006	**0.979±0.001**	0.942±0.002	0.530±0.037	0.635±0.019
	RegAD	0.696±0.015	0.246±0.031	0.314±0.036	0.643±0.032	0.855±0.021	0.908±0.013	0.881±0.014	0.919±0.018	0.911±0.001	0.566±0.048	0.558±0.159
	CoOp	0.905±0.008	0.578±0.001	0.514±0.003	0.848±0.020	0.933±0.007	0.927±0.007	0.965±0.018	0.937±0.004	0.920±0.003	0.610±0.001	0.376±0.003
	WinCLIP	0.897±0.007	**0.865±0.001**	**0.562±0.024**	**0.880±0.021**	**0.973±0.009**	**0.991±0.000**	**0.975±0.003**	0.974±0.001	**0.990±0.000**	**0.616±0.006**	**0.887±0.006**
	Ours (`InCTRL`)	**0.926±0.006**	**0.925±0.011**	**0.561±0.034**	**0.904±0.025**	**0.977±0.006**	**0.996±0.003**	**0.985±0.005**	**0.989±0.001**	**0.994±0.001**	**0.622±0.008**	**0.912±0.005**

Table 2. AUPRC results(mean±std) on nine real-world AD datasets under various few-shot AD settings. Best results and the second-best results are respectively highlighted in **red** and **blue**. 'Baseline' is a WinCLIP-based zero-shot AD model.

T	P	I	Industrial Defects				Medical Anomalies		Semantic Anomalies			
									One-vs-all		Multi-class	
			ELPV	SDD	AITEX	VisA	BrainMRI	HeadCT	MNIST	CIFAR-10	MNIST	CIFAR-10
✓	×	×	0.733	0.946	0.733	0.787	0.926	0.900	0.678	**0.924**	0.620	0.900
×	✓	×	0.794	0.946	0.730	**0.843**	0.859	0.829	**0.890**	0.850	0.504	0.894
×	×	✓	0.791	0.931	**0.796**	0.808	0.898	0.906	0.694	0.704	0.612	0.712
✓	✓	×	0.796	**0.947**	0.711	0.840	0.938	0.919	0.887	**0.924**	0.622	**0.920**
✓	×	✓	0.783	0.938	0.756	0.792	0.932	**0.924**	0.601	0.800	**0.627**	0.905
×	✓	✓	**0.816**	0.945	**0.785**	**0.856**	0.952	0.918	0.760	0.765	0.618	0.821
✓	✓	✓	**0.839**	**0.972**	0.761	**0.856**	**0.973**	**0.929**	**0.892**	**0.935**	**0.635**	**0.924**
concatenation			0.809	0.951	0.708	0.832	**0.955**	0.913	0.819	0.863	0.597	0.823
average			0.793	0.940	0.720	0.809	0.941	0.922	0.776	0.881	0.567	0.834

Table 3. AUROC results for ablation study under two-shot setting. Best results and the second-best results are respectively in **red** and **blue**. The results for VisA, and the one-vs-all settings of MNIST and CIFAR-10 represent an average result across their respective data subsets

public pre-trained backbone ViT-B/16+ in our experiments. Adam is used as the optimizer and the initial learning rate is set to 1e-3 by default. The text prompts used in `InCTRL` are kept exactly the same as WinCLIP. To enable the model to recognize both normal and abnormal objects while preventing overfitting, the training epochs is set to 10 with a batch size of 48 on a single GPU (NVIDIA GeForce RTX 3090). SPADE, PaDiM and WinCLIP[1] use the same image prompts as `InCTRL` for fair comparison, and the official implementation of PatchCore, RegAD and CoOp is taken. Further details are provided in `Appendix B`.

4.2. Main Results

Tables 1 and 2 compare the results of `InCTRL` to six SotA competing methods in AUROC and AUPRC, respectively, on nine real-world AD datasets. Note that the results for MVTec AD, VisA, and the one-vs-all settings of MNIST and CIFAR-10 represent an average result across their respective data subsets (see `Appendix C` for breakdown results). Below we analyze these results in detail.

Generalization to Industrial Defects. For the five industrial defect AD datasets, `InCTRL` significantly outperforms all competing models on almost all cases across the three few-shot settings. With more few-shot image prompts, the performance of all methods generally gets better. Specifically, Patchcore shows better performance than SPADE, PaDiM and RegAD, but all of which generalize badly on these datasets. WinCLIP obtains fairly good generalization and surpasses Patchcore, owing to CLIP's superior recognition ability. Due to the in-context residual information is well transferable across the datasets, `InCTRL` exhibits superior performance, outperforming the SotA models by a large margin, particularly on challenging datasets like ELPV and SDD. As a result, `InCTRL` model respectively gains up to 11.3%, 6.5%, 3.7% AUROC and 6.4%, 5.6%, 6% AUPRC enhancements than the best competing method.

Generalization to Medical Image Anomalies. When applied to medical image AD datasets, `InCTRL` consistently outperforms SotA models in all few-shot settings. It is evident that all competing methods perform poorly except WinCLIP. Impressively, using only two normal image prompts, `InCTRL` can obtain over 97.3% in AUPRC on BrainMRI, despite it does not have any training on medical data. On average, `InCTRL` surpass the best competing model by 3.9%, 3.4%, 3.9% in AUROC and 0.6%, 1%, 1% in AUPRC for $K = 2, 4, 8$ settings, respectively.

Generalization to Semantic Anomalies under Both One-

[1]No official implementation of WinCLIP is available. Our implementation is available at https://github.com/mala-lab/WinCLIP.

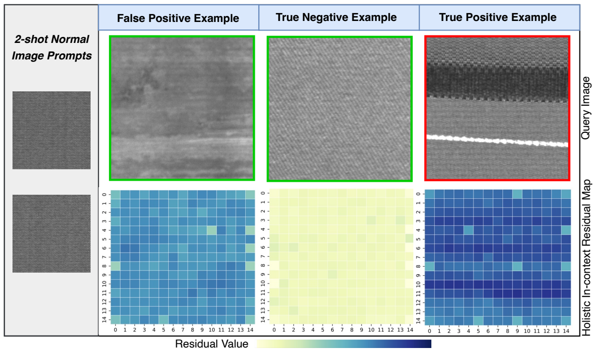

Figure 3. Visualization of query images x_t and their holistic in-context residual maps $\mathbf{M}^{+}_{x_t}$. Green and Red frames indicate normal and abnormal images respectively. Deeper colors in the residual maps represent larger residual values.

vs-all and Multi-class Settings. On detecting semantic anomalies, InCTRL again consistently surpasses all SotA models. Remarkably, InCTRL can obtain 90+% in AUROC when the previous SotA methods can obtain 50%-65% AUROC only, showcasing highly promising GAD performance. Notably, WinCLIP achieves good performance on CIFAR-10. In contrast, CoOp experiences a notable decline, presumably losing crucial semantic knowledge when adapting to the auxiliary data that diverges significantly from the semantic AD task. Overall, our InCTRL achieves the best performance with up to 8.2%, 5.1%, 4.4% AUROC and 2.3%, 1.9%, 2.5% AUPRC improvement compared to the best contender on $K = 2, 4, 8$ settings, respectively.

4.3. Why Does InCTRL Generalize Well?

Ablation Study. We examine the contribution of three key components of InCTRL on the generalization: text prompt-guided features (T), patch-level residuals (P), and image-level residuals (I), as well as their combinations. The results are reported in Table 3. The experiment results indicate that for industrial defect AD datasets, visual residual features play a more significant role compared to text prompt-based features, particularly on datasets like ELPV [18], SDD [47], and AITEX [46]. On the medical image AD datasets, both visual residuals and textual knowledge contribute substantially to performance enhancement, exhibiting a complementary relation. On semantic AD datasets, the results are dominantly influenced by patch-level residuals and/or text prompt-based features. Importantly, our three components are generally mutually complementary, resulting in the superior detection generalization across the datasets.

Significance of In-context Residual Learning. To assess the importance of learning the residuals in InCTRL, we experiment with two alternative operations in both multi-layer patch-level and image-level residual learning: replacing the residual operation with 1) a **concatenation** operation and 2) an **average** operation, with all the other components of InCTRL fixed. As shown in Table 3, the in-context residual learning generalizes much better than the other two alternative ways, significantly enhancing the model's performance in GAD across three distinct domains.

4.4. Failure Cases

To understand the results of InCTRL better, we provide visualization results illustrating both successful detection and failures by InCTRL on ELPV [18]. As depicted in Figure 3, the incorrectly identified anomalies (False Positive example) shows substantially texture difference compared to the two normal image prompts, leading to similarly large residual values as that for a True Positive example. In contrast, when the query image resembles the normal image prompts well, the residual values are clearly very small, as shown by the True Negative example. These cases may be remedied when the image prompts include similar normal images as the falsely identified anomaly. This failure case exemplifies the challenge of GAD using only a few image prompts.

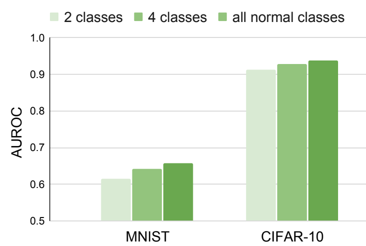

Figure 4. AUROC of InCTRL with sample prompts from varying numbers of normal classes.

This case is related to the diversity of few-shot normal sample prompts, which is particularly important when the normal data is complex, *e.g.*, having multiple different normal patterns in our multi-class protocol.

To rigorously investigate this problem, we evaluate the performance of the two datasets under the multi-class protocol, with varying numbers of normal classes included in the normal prompts in the eight-shot setting. As shown in Figure 4, the performance of InCTRL continually improves when the prompt sets have samples from more normal classes.

5. Conclusion

In this work we introduce a GAD task to evaluate the generalization capability of AD methods in identifying anomalies across various scenarios without any training on the target datasets. This is the first study dedicated to a generalist approach to anomaly detection, encompassing industrial defects, medical anomalies, and semantic anomalies. Then we propose an approach, called InCTRL, to addressing this problem under a few-shot setting. InCTRL achieves a superior GAD generalization by holistic in-context residual learning. Extensive experiments are performed on nine AD datasets to establish a GAD evaluation benchmark for the aforementioned three popular AD tasks, on which InCTRL significantly and consistently outperforms SotA competing models across multiple few-shot settings.

References

[1] Abhishek Aich, Kuan-Chuan Peng, and Amit K Roy-Chowdhury. Cross-domain video anomaly detection without target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2579–2591, 2023. 2

[2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019. 1, 2

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 3

[5] Niamh Belton, Misgina Tsighe Hagos, Aonghus Lawlor, and Kathleen M Curran. Fewsome: One-class few shot anomaly detection with siamese networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2977–2986, 2023. 2, 3

[6] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020. 1, 2

[7] 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. 6

[8] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 1, 2

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[10] Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. *Medical Image Analysis*, 86:102794, 2023. 2

[11] Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6511–6523, 2023. 1, 2, 6

[12] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024. 1

[13] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 3

[14] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 3

[15] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 383–392, 2022. 1, 2

[16] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 2, 3, 4, 6

[17] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 2, 3, 4, 6

[18] Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, 185:455–468, 2019. 6, 8

[19] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 1, 2

[20] 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. 2

[21] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 3

[22] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8791–8800, 2021. 1, 2

[23] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pages 303–319. Springer, 2022. 2, 3, 6

[24] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. Openclip. *Zenodo*, 4:5, 2021. 6

[25] 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. 2, 3, 5, 6

[26] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim:

A unified modeling approach for vision with learned guiding codes. *Advances in Neural Information Processing Systems*, 35:26295–26308, 2022. 3

[27] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009. *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5, 2009. 6

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[29] Jingyi Liao, Xun Xu, Manh Cuong Nguyen, Adam Goodge, and Chuan Sheng Foo. Coft-ad: Contrastive fine-tuning for few-shot anomaly detection. *arXiv preprint arXiv:2402.18998*, 2024. 3

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[31] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12147–12156, 2023. 1, 2

[32] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3

[33] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 125–141. Springer, 2020. 2

[34] Guansong Pang, Kai Ming Ting, and David Albrecht. Lesinn: Detecting anomalies by identifying least similar nearest neighbours. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pages 623–630. IEEE, 2015. 4

[35] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2041–2050, 2018. 4

[36] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021. 1

[37] Guansong Pang, Chunhua Shen, Huidong Jin, and Anton van den Hengel. Deep weakly-supervised anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1795–1807, 2023. 6

[38] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 1, 2

[39] 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. 2, 4

[40] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 2, 3, 4, 6

[41] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2020. 1, 2, 6

[42] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 1, 2, 6

[43] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019. 1, 2

[44] Eli Schwartz, Assaf Arbelle, Leonid Karlinsky, Sivan Harary, Florian Scheidegger, Sivan Doveh, and Raja Giryes. Maeday: Mae for few and zero shot anomaly-detection. *arXiv preprint arXiv:2211.14307*, 2022. 2, 3

[45] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8495–8504, 2021. 2, 3

[46] Javier Silvestre-Blanes, Teresa Albero-Albero, Ignacio Miralles, Rubén Pérez-Llorens, and Jorge Moreno. A public fabric database for defect detection methods and results. *Autex Research Journal*, 19(4):363–374, 2019. 6, 8

[47] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020. 6, 8

[48] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004. 1, 2

[49] Yu Tian, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 128–140. Springer, 2021. 2

[50] Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. *Medical image analysis*, 90:102930, 2023. 2

[51] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and

Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24511–24520, 2023. 1, 2

[52] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021. 1, 2

[53] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3

[54] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 3

[55] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 3

[56] Ze Wang, Yipin Zhou, Rui Wang, Tsung-Yu Lin, Ashish Shah, and Ser Nam Lim. Few-shot fast-adaptive anomaly detection. *Advances in Neural Information Processing Systems*, 35:4957–4970, 2022. 2, 3

[57] Jhih-Ciang Wu, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Learning unsupervised metaformer for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4369–4378, 2021. 2, 3

[58] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. *arXiv preprint arXiv:2311.07042*, 2023. 2

[59] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *arXiv preprint arXiv:2308.11681*, 2023. 2

[60] Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Squid: Deep feature in-painting for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23890–23901, 2023. 1, 2

[61] Guoyang Xie, Jingbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. *arXiv preprint arXiv:2301.12082*, 2023. 2, 3

[62] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023. 1

[63] Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3110–3118, 2021. 1, 2

[64] Xincheng Yao, Ruoqi Li, Zefeng Qian, Yan Luo, and Chongyang Zhang. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6803–6813, 2023.

[65] Xincheng Yao, Chongyang Zhang, Ruoqi Li, Jun Sun, and Zhenyu Liu. One-for-all: Proposal masked cross-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4792–4800, 2023. 1, 2

[66] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 2

[67] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020. 1, 2

[68] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.

[69] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 1, 2

[70] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection, 2023. 1, 2

[71] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 6

[72] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[73] Jiawen Zhu, Choubo Ding, Yu Tian, and Guansong Pang. Anomaly heterogeneity learning for open-set supervised anomaly detection. *arXiv preprint arXiv:2310.12790*, 2023. 2

[74] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 6