Few-shot Anomaly Detection via Personalization

Sangkyung Kwak [1] **Jongheon Jeong** [1] **Hankook Lee** [2] **Woohyuck Kim** [1] **Jinwoo Shin** [1]

Abstract

Even with a plenty amount of normal samples, anomaly detection has been considered as a challenging machine learning task due to its one-class nature, *i.e.*, the lack of anomalous samples in training time. It is only recently that a *few-shot* regime of anomaly detection became feasible in this regard, *e.g.*, with a help from large vision-language pre-trained models such as CLIP, despite its wide applicability. In this paper, we explore the potential of large *text-to-image generative models* in performing few-shot anomaly detection. Specifically, recent text-to-image models have shown unprecedented ability to generalize from few images to extract their common and unique concepts, and even encode them into a textual token to "personalize" the model: so-called *textual inversion*. Here, we question whether this personalization is specific enough to discriminate the given images from their potential anomalies, which are often, *e.g.*, open-ended, local, and hard-to-detect. We observe that the standard textual inversion is not enough for detecting anomalies accurately, and thus we propose a simple-yet an effective regularization scheme to enhance its specificity derived from the zero-shot transferability of CLIP. We also propose a self-tuning scheme to further optimize the performance of our detection pipeline, leveraging synthetic data generated from the personalized generative model. Our experiments show that the proposed inversion scheme could achieve state-of-the-art results on a wide range of few-shot anomaly detection benchmarks.

1. Introduction

The ability to identify unusual patterns in images is a natural capability of human cognition. Even when provided with only a small number of normal examples, humans can adapt to discriminate abnormality from the examples, whereas this remains as a challenging task in the field of computer vision. *Anomaly detection* (AD), where the task is formulated, faces fundamental challenges due to several reasons. Firstly, objects and their defects can vary widely in terms of color, texture, and size across numerous industrial domains: *e.g.*, aerospace, automobiles, pharmaceuticals, and electronics. Besides, some types of anomaly can be fine-grained which has only little differences between normal and anomalous data while other can be coarse-grained. Secondly, obtaining and specifying the expected variations in defects is limited and costly in real-world situations.

Upon these fundamental challenges, significant efforts have been made to approach AD: *e.g.*, either in *one-class*, unsupervised setting (Bergmann et al., 2022; Cohen & Hoshen, 2020; Defard et al., 2021; Li et al., 2021; Ristea et al., 2022; Roth et al., 2022; Zavrtanik et al., 2021), or in semi-supervised setting (Zou et al., 2022), to name a few. Intuitively, the major technical bottleneck here is to learn features expressive enough to encode inter-class variability without knowing anomalous data, while maintaining intra-class variability induced from normal data. To overcome this, there have been two representative approaches: (a) *feature-based approaches* (Ruff et al., 2018; Yi & Yoon, 2020; Defard et al., 2021; Roth et al., 2022; Cohen & Hoshen, 2020; Deng & Li, 2022; Gudovskiy et al., 2022) leverage an external, pre-trained feature extractor, *e.g.*, on ImageNet, to retrieve its richer features in modeling AD; and (b) *reconstruction-based approaches* (Akcay et al., 2018; Zavrtanik et al., 2021; Ristea et al., 2022; Baur et al., 2019; Gong et al., 2019) instead model a generative model to extract faithful features in the normal data available, in an attempt to improve the sensitivity of features. With hundreds to thousands of normal images, such approaches have shown effectiveness to achieve high-enough detection performances, *e.g.*, on existing industrial anomaly detection benchmarks (Bergmann et al., 2019; Zou et al., 2022).

Anomaly detection with limited data, *e.g.*, with only *few normal images*, has been still challenging even until recently. The cost-efficiency of a *language-driven prior* has emerged as an effective way to mitigate the challenge, particularly since CLIP (Radford et al., 2021), a recent large vision-language model. For example, Jeong et al. (2023)

[1]KAIST [2]LG AI Research. Correspondence to: Jinwoo Shin <jinwoos@kaist.ac.kr>.

2^{nd} *AdvML Frontiers workshop at* 40^{th} *International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023.

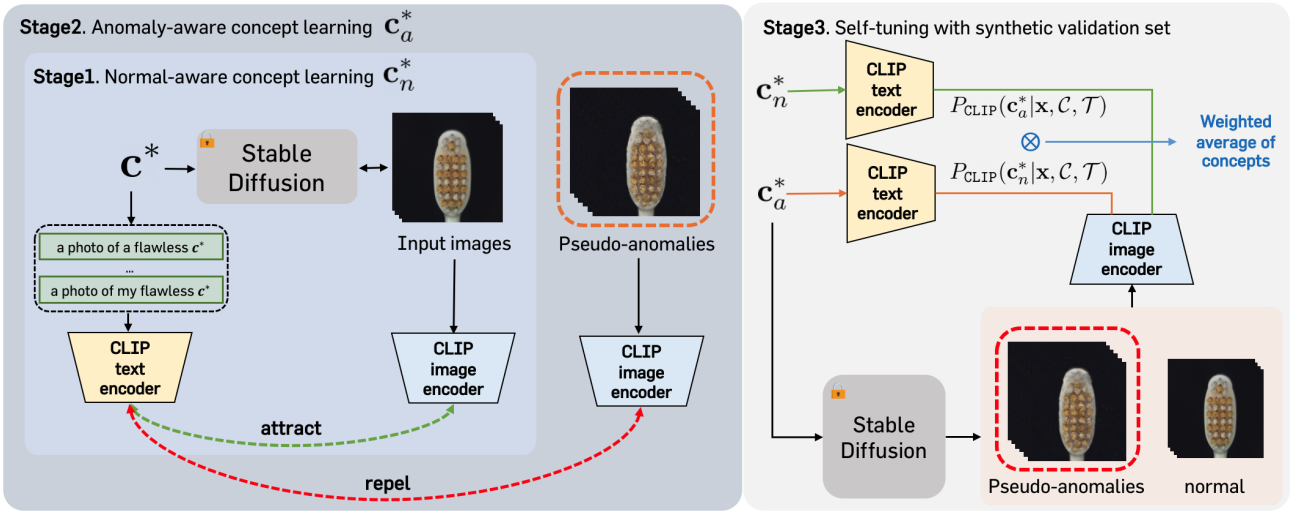

Figure 1. Overview of *Anomaly Detection via Personalization* (ADP). In Stage1, normal concepts are converted into $\mathbf{c}_n^*$ by guiding the normal prompt incorporating $\mathbf{c}_n^*$ to be closer to the given images (Section 3.1). In Stage2, anomalous concepts are converted into $\mathbf{c}_a^*$ by additionally distancing pseudo-anomalies to the normal prompt $\mathbf{c}_a^*$ incorporating (Section 3.2). In Stage3, by utilizing CLIP, the use of $\mathbf{c}_n^*$ and $\mathbf{c}_n^*$ are further tuned with synthesized pseudo-anomalies (Section 3.3).

have demonstrated state-of-the-art performances in few-shot AD by incorporating a "zero-shot", language-driven AD pipeline from CLIP, *e.g.*, by additionally comparing similarities to words "normal" *vs.* "damaged" for a given image: a similar exploration has been made in the context of novelty detection (or so-called out-of-distribution detection) by Ming et al. (2022).

Although it is evident that language can be a useful prior for AD, *e.g.*, to clarify the vague concepts of abnormality by supplying label words (*e.g.*, bottle, capsule, *etc.*), the current interface of "hand-crafting" language prompts becomes a limiting bottleneck as the given AD task gets more specific to the (few-shot) data: and accordingly as it gets "harder-to-describe". In turn, it is observed that the performance of current language-based AD is highly dependent by the prompt design, which is heuristic in nature and requires a careful tuning by humans. For example, Jeong et al. (2023) indeed assumed the knowledge of class labels as text words in performing their zero-/few-shot AD.

Contribution. In this paper, we propose a new design of language-based AD, coined *Anomaly Detection via Personalization* (ADP), which leverages *model personalization* (Gal et al., 2022; Ruiz et al., 2022; Kumari et al., 2022) that is recently enabled by large-scale text-to-image generative models (Rombach et al., 2021; Saharia et al., 2022). Specifically, recent text-to-image generative models have shown capabilities to extract detailed concepts shared across a few given images, and encode them as a *textual token* to compose natural language sentences associated with the generative model: it can "personalize" the model to generate images containing the concepts. Here, we focus on exploring whether this new ability of *textual inversion* could replace the current brittleness in crafting few-shot, language-based AD in practice. We first observe that the current objective for textual inversion (in the context of generative modeling) may not be specific enough to perform accurate few-shot AD. Motivated by this, we propose a novel textual inversion scheme to improve its specificity, based on a richer guidance induced by CLIP (Radford et al., 2021). We develop a two-stage inversion scheme designed for general AD: the former to personalize from normal samples, and the latter to refine itself based on the personalized model, particularly leveraging the "synthetic" anomaly samples that the model can generate. In this way, the inversion can better capture fine-grained visual semantics which is demanded to perform an accurate AD. We also propose to re-utilize the anomaly synthesis scheme for a self-tuning of our AD model, which is a unique ability to our framework.

With the proposed method, we tackle *extreme few-normal-shot* AD, *viz.*, 2 to 16, an under-explored setup due to its difficulty (Rudolph et al., 2021; Sheynin et al., 2021; Huang et al., 2022). We summarize our main contributions in what follows:

- We introduce a novel method to capture unique concepts of anomalies into the token, which improves few-shot AD.
- Using the anomaly-aware token, we show that we can effectively synthesize pseudo-anomalies with pre-trained text-to-image diffusion model.

- We propose a simple yet effective self-tuning method to utilize the tokens in the pre-trained vision-language model for AD.
- Through an extensive evaluation on MVTec-AD and VisA, we report new state-of-the-art results on few-shot AD, *e.g.*, **97.1%** on MVTec-AD and **89.7%** on VisA in AUROC in 16-shot AD, notably even without text descriptions on the object labels as assumed in prior art (Jeong et al., 2023).

2. Preliminaries

2.1. Problem setup

Anomaly detection (AD) aims to determine the presence of "abnormality" given an image $\mathbf{x} \in \mathcal{X}$. We formulate AD as a binary classification problem $\mathcal{X} \to \{0, 1\}$, where "1" indicates the presence of abnormality. Due to the lack of anomalous samples in practice, AD is often assumed to be *one-class*, *i.e.*, its training data $\mathcal{D} := \{(x_i, 0)\}_{i=1}^K$ consists of only normal (or negative) samples. In this work, we follow this one-class protocol, particularly focusing on *extreme few-shot* scenarios where the training data only consists of $K = 2$ to 16 normal images. It is also a practice to cast AD as a problem of assigning *anomaly score* rather than a direct classification, again due to the high-imbalance in data: the actual classification in practice is done by thresholding the score.

To solve this extreme few-shot AD task, we utilize vision-language foundation models, a contrastive encoder (*e.g.*, CLIP (Radford et al., 2021)) and a diffusion model (*e.g.*, LDMs (Rombach et al., 2021)), pre-trained on external datasets. Our approach is widely applicable as the foundation models have shown to be generalizable across various downstream tasks and they are publicly available. We will describe the vision-language contrastive encoder and diffusion model in Section 2.2 and Section 2.3, respectively.

2.2. Contrastive language image pre-training

Contrastive language image pre-training (CLIP) (Radford et al., 2021) is a large-scale pre-training method that offers a joint vision-language representation by training an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$ using contrastive learning (Chen et al., 2020; Zhang et al., 2020) with the million-scale image-text pairs from the web. One attractive ability of CLIP is zero-shot transfer, especially for image classification. To be specific, given a set of labels $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_N\}$, an image $\mathbf{x}$ can be classified by the following probability:

$$P_{\texttt{CLIP}}(\mathbf{c}_i|\mathbf{x}, \mathcal{C}, \mathcal{T}) := \frac{\exp\left(\texttt{sim}\left(f(\mathbf{x}), \mathcal{G}(\mathbf{c}_i)\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\texttt{sim}\left(f(\mathbf{x}), \mathcal{G}(\mathbf{c}_j)\right)/\tau\right)}, \tag{1}$$

where $\mathcal{G}(\mathbf{c}_i) = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} g(T(\mathbf{c}_i))$, $\texttt{sim}(\cdot,\cdot)$ is the cosine similarity, $\tau > 0$ is the temperature hyperparameter, and $T \in \mathcal{T}$ is a prompt template attached to a label $\mathbf{c}$ such as "`a photo of a [c]`". Note that using multiple templates, *i.e.*, template ensemble, can improve the zero-shot classification accuracy (Radford et al., 2021).

2.3. Textual inversion

Textual inversion (Gal et al., 2022) aims to extract the common concept $\mathbf{c}$ of images into a text embedding and use it for a *personalized* generation, *i.e.*, to sample from the distribution $p(\mathbf{x}|\mathbf{c})$. To this end, we use a pre-trained text-to-image latent diffusion model (Rombach et al., 2021), $p_{\texttt{t2i}}(\mathbf{x}|\mathbf{s})$, where $\mathbf{s}$ is a conditioning text. Given a set of images $\{\mathbf{x}_i\}_{i=1}^K$, the textual inversion finds their common concept $\mathbf{c}^*$ by solving the following optimization problem:

$$\mathbf{c}^* := \arg\max_{\mathbf{c}} \sum_{i=1}^K \sum_{T\in\mathcal{T}} \log p_{\texttt{t2i}}(\mathbf{x}_i|T(\mathbf{c})), \tag{2}$$

where $\mathcal{T}$ is a set of prompt templates. After textual inversion, one can generate a new image of the concept $\mathbf{c}^*$ with a template $T \in \mathcal{T}$, *i.e.*, $\mathbf{x} \sim p_{\texttt{t2i}}(\cdot|T(\mathbf{c}^*))$. Gal et al. (2022) found that the concept is well-optimized with only a few images, *e.g.*, $K = 4$. In addition, since the model uses the CLIP text encoder g for text conditioning, one can use the concept on the CLIP representation space.

3. Anomaly detection via personalization

In this section, we introduce Anomaly Detection via Personalization (ADP), a novel framework for few-shot anomaly detection utilizing the ground knowledge in vision-language foundation models. To be specific, ADP finds the concept that can (i) generate both normal and abnormal images via textual inversion and also (ii) detect the abnormality via CLIP zero-shot classification. ADP then performs anomaly detection using the concept and multi-level image features.

To perform anomaly detection (*i.e.*, 2-way classification) without label information, we utilize normal and anomalous state templates, S_n and S_a, respectively, for a concept $\mathbf{c}$, following Jeong et al. (2023):

$$S_n(\mathbf{c}) := \text{"}\texttt{flawless}\ [\mathbf{c}]\text{"}, \quad S_a(\mathbf{c}) := \text{"}\texttt{damaged}\ [\mathbf{c}]\text{"}.$$

Given a set of a few normal images $\{\mathbf{x}_i\}_{i=1}^K$, the state templates S_n and S_a, and a set of prompt templates $\mathcal{T}$ (*e.g.*, "`a photo of [`$\mathbf{c}_n$`]`"), ADP follows the following procedure:

Step 1. Find the *normal-aware concept* $\mathbf{c}_n^*$ by guiding *normal* images to be close to the *normal* state prompts (Section 3.1).

Step 2. Find the *anomaly-aware concept* $\mathbf{c}_a^*$ by guiding *pseudo-anomalous* images to put distance to the *normal* state prompts (Section 3.2).

Step 3. Perform anomaly detection using the concepts, $\mathbf{c}_n^*$ and $\mathbf{c}_a^*$ (Section 3.3).

3.1. Normal-aware concept learning

In normal-aware concept learning, we aim to capture the visual normal concept $S_n(\mathbf{c}_n^*)$ from the normal images $\{\mathbf{x}_i\}_{i=1}^K$. To this end, in addition to textual inversion, we make CLIP embeddings of the images similar with that of the normal state prompt $T(S_n(\mathbf{c}))$, *e.g.*, "`a photo of a flawless [c]`", while dissimilar with that of the anomalous state prompt $T(S_a(\mathbf{c}))$, *e.g.*, "`a rendering of a damaged [c]`". Formally, the normal-aware concept $\mathbf{c}_n^*$ can be obtained by solving the following optimization problem:

$$\begin{aligned}\mathbf{c}_n^* &= \arg\max_{\mathbf{c}} \mathcal{J}_n(\mathbf{c}; \{\mathbf{x}_i\}_{i=1}^K) \\ &:= \sum_{i=1}^{K} \sum_{T\in\mathcal{T}} \log p_{\texttt{t2i}}(\mathbf{x}_i|T(\mathbf{c})) \\ &\qquad + \alpha P_{\texttt{CLIP}}(S_n(\mathbf{c})|\mathbf{x}_i, \mathcal{C}(\mathbf{c}), \mathcal{T}) \quad (3)\end{aligned}$$

where $\mathcal{C}(\mathbf{c}) = \{S_n(\mathbf{c}), S_a(\mathbf{c})\}$ is the set of normal and anomalous labels of the concept $\mathbf{c}$ and α is a hyperparameter. We initialize the concept $\mathbf{c}$ as the word "`object`" which is applicable to regardless of the domain and dataset.

3.2. Anomaly-aware concept learning

We here aim to further capture the "anomalous" concept $S_a(\mathbf{c}_a^*)$ from anomalous images while maintaining the normal concept $S_n(\mathbf{c}_a^*)$ of the normal images $\{\mathbf{x}_i\}_{i=1}^K$. To this end, we first synthesize *pseudo-anomalous* images via text-guided image manipulation (Meng et al., 2021) using the text-to-image diffusion model $p_{\texttt{t2i}}(\mathbf{x}|\mathbf{s})$. We here use a normal image $\mathbf{x}_i$ as a reference image and "`a photo with damage`" or "`a photo of an object with damage`" as a conditioning text $\mathbf{s}$. To give more diversity, the manipulated images are further augmented with random resizing and cropping. We denote $\{\tilde{\mathbf{x}}_j\}_{j=1}^L$ as synthesized pseudo-anomalous images (*i.e.*, *pseudo-anomalies*). The examples of pseudo-anomalies are illustrated in the supplementary.

In addition to normal-aware concept learning, we put distance between the pseudo-anomalous images $\{\tilde{\mathbf{x}}_j\}_{j=1}^L$ and the normal state prompt $S_n(\mathbf{c}_a^*)$. Formally, the anomaly-aware concept $\mathbf{c}_a^*$ can be obtained by solving the following optimization problem:

$$\begin{aligned}\mathbf{c}_a^* &= \arg\max_{\mathbf{c}} \mathcal{J}_a(\mathbf{c}; \{\mathbf{x}_i\}, \{\tilde{\mathbf{x}}_j\}) \\ &:= \mathcal{J}_n(\mathbf{c}; \{\mathbf{x}_i\}) \\ &\qquad - \alpha \sum_{j=1}^{L} \left(P_{\texttt{CLIP}}(S_n(\mathbf{c})|\tilde{\mathbf{x}}_j, \mathcal{C}(\mathbf{c}), \mathcal{T}) - \gamma\right)^+, \quad (4)\end{aligned}$$

where $(\cdot)^+ := \max(\cdot, 0)$, $\mathcal{C}(\mathbf{c}) = \{S_n(\mathbf{c}), S_a(\mathbf{c})\}$, α and γ are hyperparameters. We initialize the concept $\mathbf{c}$ as the normal-aware concept $\mathbf{c}_n^*$ described in Section 3.1. Since $\mathbf{c}_n^*$ captures high-level visual features of normal images, initializing the concept with the normal-aware helps learning fine-grained anomalous features.

3.3. Anomaly detection with learned concepts

We now introduce a simple yet effective detection scheme using the learned concepts. At a high-level, our scheme first extracts CLIP text embeddings of all available prompt state templates with the concepts and then mix them to construct effective 2-way classification prototypes via *self-tuning*. Given a test image, we detect whether it is in-distribution or not using its CLIP image embedding.

Self-tuning. To utilize both concepts effectively, we mix the concepts using importance weights obtained by a pseudo-validation set, which consists of the normal images $\{\mathbf{x}_i\}_{i=1}^K$ and new pseudo-anomalous images $\{\tilde{\mathbf{x}}_j\}_{j=1}^{L'}$ synthesized by conditioning texts, "`a photo of a damaged [`$\mathbf{c}_a^*$`]`" and "`a photo of a [`$\mathbf{c}_a^*$`] with damage`", as described in Section 3.2. The importance weight $w(\mathbf{c})$ of each concept $\mathbf{c}$ can be computed by evaluating CLIP zero-shot classification as follows:

$$\begin{aligned}w(\mathbf{c}) := \frac{1}{K}\sum_{i=1}^{K} P_{\texttt{CLIP}}(S_n(\mathbf{c})|\mathbf{x}_i, \mathcal{C}(\mathbf{c}), \mathcal{T}) \\ + \frac{1}{L'}\sum_{j=1}^{L'} P_{\texttt{CLIP}}(S_a(\mathbf{c})|\tilde{\mathbf{x}}_j, \mathcal{C}(\mathbf{c}), \mathcal{T}). \quad (5)\end{aligned}$$

We then compute the weighted average of the CLIP text embeddings to construct the classification prototype vectors using the CLIP text encoder g as follows:

$$\mathbf{p}_{\mathbf{s}} := \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \frac{w(\mathbf{c}_n^*)\cdot g(T(S_{\mathbf{s}}(\mathbf{c}_n^*))) + w(\mathbf{c}_a^*)\cdot g(T(S_{\mathbf{s}}(\mathbf{c}_a^*)))}{w(\mathbf{c}_n^*) + w(\mathbf{c}_a^*)}. \quad (6)$$

Anomaly detection. Given a test image $\mathbf{x}$, our detection score $\texttt{ADP}(\mathbf{x})$ is formally defined by

$$\texttt{ADP}(\mathbf{x}) := \frac{\exp\left(\texttt{sim}(f(\mathbf{x}), \mathbf{p}_n)/\tau\right)}{\exp\left(\texttt{sim}(f(\mathbf{x}), \mathbf{p}_n)/\tau\right) + \exp\left(\texttt{sim}(f(\mathbf{x}), \mathbf{p}_a)/\tau\right)}. \quad (7)$$

To further improve detection performance, we utilize visual features (*i.e.*, feature maps) using the CLIP image encoder as Jeong et al. (2023). Specifically, we add the feature similarity score computed by patch-wise spatial feature association between the test image $\mathbf{x}$ and the given normal images $\{\mathbf{x}_i\}_{i=1}^K$ into our score $\texttt{ADP}(\mathbf{x})$.

4. Experiments

We conduct an extensive evaluation on the proposed method, ADP, on MVTec-AD (Bergmann et al., 2019) and VisA (Zou et al., 2022) benckmarks, two popular datasets in AD capturing real-world scenarios of industrial defect detection. In particular, we mainly evaluate under extreme few-shot regimes, *i.e.*, by assuming K-shot of normal images for each task. The detailed experimental setups, *e.g.*, hyperparameters, preprocessing, are provided in the supplementary.

Implementation details. Throughout our experiments, we use Stable Diffusion v2-1[1] as the backbone text-to-image model, which uses the CLIP text encoder for conditioning: so that compatible with our framework which utilizes CLIP as well. We use the OpenCLIP implementation[2] of CLIP ViT-H/14 model trained on the LAION-2B English subset of LAION-5B, following the choice of the Stable Diffusion v2-1 model we are based on. We use our re-implementation of WinCLIP (Jeong et al., 2023) for our experiments, which we have confirmed the reproducibility of the results.

4.1. Results

For each setup, we report two versions of our method: (a) **ADP**, the default version introduced in (7) that does *not* relying on specific label texts (*e.g.*, "transistor") as considered in WinCLIP+ (Jeong et al., 2023); in addition, we also report (b) **ADP**$_\ell$ which also incorporate the knowledge of label texts by weight-averaging also the class label ℓ at (6) as well as $\mathbf{c}_n^*$ and $\mathbf{c}_a^*$ using $w(\ell)$ as the weight. We use *Area Under Receiver Operator Characteristic-curve* (AUROC) as the major evaluation metric. We report our results with standard deviation across 3 different random seeds.

In Table 1, we report the overall performances of our methods, "ADP" and "ADP$_\ell$", for 2-, 4-, 8- and 16- shots AD compared to the baseline on MVTec-AD and VisA:[3] ADP and ADP$_\ell$ significantly outperform the state-of-the-art results of WinCLIP+ on both datasets. On both MVTec-AD and VisA, we observe that our approach of ADP exhibits a wider performance gap over WinCLIP+ as more shots are given: specifically, on the 4-shot MVTec-AD, ADP outperforms WinCLIP+ by 1.7% in AUROC, while it does by 2.3% on the 16-shot scenario. Similarly, in the case of VisA, ADP improves over WinCLIP+ by 3.1% in AUROC on 4-shot, while it does by 4.7% on the 16-shot setup. Regarding the performance of ADP$_\ell$ over ADP: although the knowledge of label texts in ADP$_\ell$ does helpful to improve our results on low-shot setups, *e.g.*, 4-shot, we observe that ADP gradually matches the performance with ADP$_\ell$ having with more shots: in the 16-shot scenario of MVTec-AD, ADP even shows a consistently better performances over ADP$_\ell$ when viewed in class-wise, achieving 97.1% in AUROC.

[1] https://github.com/Stability-AI/stablediffusion

[2] https://github.com/openai/CLIP

[3] We report the detailed results of Table 1 in the supplementary.

4.2. Ablation study

Comparison with textual inversion. In Table 2, we compare our proposed concepts with the standard textual inversion (Gal et al., 2022) in the context of AD. Specifically, we compare our results on 4-shot MVTec-AD and VisA with an ablation that the steps for concept optimization are replaced by the standard version of textual inversion (as reported by "TI" in Table 2). We examine the results obtained by incorporating only $\mathbf{c}_n^*$ and $\mathbf{c}_a^*$ in the text prompts, as well as by incorporating both concepts $\mathbf{c}_n^* + \mathbf{c}_a^*$, denoted as ADP. Overall, we observe that converting only via textual inversion, *e.g.*, encoding tokens simply through the reconstruction loss, falls short specifically in few-shot AD. For example, on the VisA dataset we observe that ADP improves upon the original textual inversion by 9.2% in AUROC. These results highlight the suitability of ADP to effectively capture the concepts related to *abnormality* into tokens through an additional guidance via CLIP.

In terms of learned concepts, mixing $\mathbf{c}_n^*$ and $\mathbf{c}_a^*$ via ADP (Section 3.3) leads to a better performance, confirming the effectiveness of our self-tuning scheme.6 We observe that the performance itself of $\mathbf{c}_a^*$ as an individual concept may not be significantly better compared to $\mathbf{c}_n^*$. A clear performance gain could be obtained by combining the two concepts, however, confirming that $\mathbf{c}_n$ and $\mathbf{c}_a$ complement each other.

Qualitative comparison. Prior studies have proposed synthesizing *anomalous* images by adding visually irregular appearances into normal images (Li et al., 2021; Yang et al., 2022; Zavrtanik et al., 2021). In this paper, we take a different approach which generates pseudo-anomalies using a pre-trained text-to-image diffusion model (Meng et al., 2021). Specifically, this is achieved by adding noise to a given reference image and conditioning the reconstruction process on text prompts. We investigate two types of prompts: (1) simple text prompts, such as "a photo with damage", (2) prompts incorporating the *anomaly-aware* concept $\mathbf{c}_a$, such as "a photo of a $\mathbf{c}_a$ with damage". The results are given in the supplementary. It demonstrates that the use of *anomaly-aware* concept $\mathbf{c}_a$ leads to the generation of fine-grained anomalies, compared to simple text prompts.

Table 1. Anomaly detection (AD) performance on MVTec-AD and VisA benchmark on 2-, 4-, 8-, and 16- shots. We report the mean AUROC (%) and standard deviation over three random seeds for each measurement, which highest AUROC (%) is marked as bold.

	2-shot			4-shot			8-shot			16-shot		
Data \ Method	WinCLIP+	**ADP**	**ADP$_\ell$**	WinCLIP+	**ADP**	**ADP$_\ell$**	WinCLIP+	**ADP**	**ADP$_\ell$**	WinCLIP+	**ADP**	**ADP$_\ell$**
MVTec-AD	93.8$_{\pm 1.0}$	94.4$_{\pm 1.2}$	**95.4**$_{\pm 0.9}$	94.1$_{\pm 0.7}$	95.8$_{\pm 1.1}$	**96.2**$_{\pm 0.8}$	94.6$_{\pm 0.1}$	96.8$_{\pm 0.4}$	**97.0**$_{\pm 0.2}$	94.8$_{\pm 0.1}$	**97.1**$_{\pm 0.5}$	97.0$_{\pm 0.3}$
VisA	84.2$_{\pm 0.2}$	85.7$_{\pm 0.9}$	**86.9**$_{\pm 0.9}$	84.6$_{\pm 0.4}$	87.7$_{\pm 0.3}$	**88.4**$_{\pm 0.4}$	85.0$_{\pm 0.0}$	88.6$_{\pm 0.3}$	**89.2**$_{\pm 0.1}$	85.0$_{\pm 0.1}$	89.7$_{\pm 0.9}$	**90.1**$_{\pm 0.5}$

Table 2. Comparison of anomaly detection (AD) with naïve textual inversion and across the use of learned concepts in MVTec-AD and ViSA dataset for 4-shot. "TI" denotes the naïve textual inversion.

Data \ Method	TI	$\mathbf{c}_n^*$	$\mathbf{c}_a^*$	**ADP**
MVTec-AD	88.6	95.9	95.2	**96.0**
VisA	78.5	87.1	85.3	**87.7**

5. Related work

Anomaly detection. In the field of anomaly detection, the focus has been on one-class methods that utilize a large amount of normal images (Li et al., 2021; Defard et al., 2021; Cohen & Hoshen, 2020; Yi & Yoon, 2020; Zavrtanik et al., 2021; Yu et al., 2021). Specifically, in industrial anomaly detection, which requires to learn unique nominal features, recent works suggest utilizing pre-trained models with external image dataset (Cohen & Hoshen, 2020; Defard et al., 2021). However, these existing approaches encounter limitations when applied to specific applications due to the challenges posed by the full-normal-shot setup in MVTec-AD benchmark (Bergmann et al., 2019). Recent studies (Rudolph et al., 2021; Sheynin et al., 2021) have investigated few-shot setups by employing augmentation techniques to expand the small support set, leading to enhanced modeling of normality. Another approach, RegAD (Huang et al., 2022), introduces the concept of model re-using which pre-trains an object-agnostic registration network with diverse images to establish normality for unseen objects. Additionally, utilizing pre-trained vision-language model to extract the prior knowledge has shown remarkable improvement in few-normal-shot anomaly detection (Jeong et al., 2023). The few-shot setups in anomaly detection is still under-explored and has room for improvement.

Text-to-image diffusion models. At a high level, diffusion models (Ho et al., 2020; Song et al., 2020), class of generative models, learn the target distribution $p_{\text{data}}(\mathbf{x})$ by learning a gradual denoising process from Gaussian prior distribution to reach $p_{\text{data}}(\mathbf{x})$. The field of diffusion models has seen a wide range of applications, including text-to-image generation. Text-to-image diffusion models are able to generate images conditioned by text prompts (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2021), which show promising result in image synthesis. Among them, one notable approach is Stable Diffusion (Rombach et al., 2021), which is a popular variant of latent diffusion models (LDMs) (Rombach et al., 2021). This is trained on extremely large-scale data, have demonstrated remarkable generalization ability. To utilize the strong generalizability in synthesizing images, we incorporate Stable Diffusion to address anomaly detection task.

Personalization of text-to-image models. With the outstanding scalability of pre-trained text-to-image diffusion models, recent works make great efforts to generate specific instances like personal animals or rare categories. To inject the new concept to the pre-trained models while preserving the previous knowledge, recent works suggest several approaches. This includes fine-tuning only subset of the parameters (Kumari et al., 2022), fine-tuning with the method to preserve prior knowledge (Ruiz et al., 2022) and introducing and optimizing a word vector for the new concept (Gal et al., 2022). In this way, models excel at integrating new information into their domain without forgetting the prior or overfitting to a small subset of training images. Motivated from this, we suggest utilizing model personalization in identifying anomalies, which enables addressing few-shot setting in anomaly detection task.

6. Conclusion

In this paper, we propose *Anomaly Detection via Personalization* (ADP), a novel approach to address the challenging problem of few-shot anomaly detection based on recent text-to-image diffusion models. We show that aligning state prompts with image features effectively guides the model to learn concepts related to *normal* and *anomalous* instances. Additionally, we introduce synthesizing pseudo-anomalies using a personalized generative model based on the learned concepts. By incorporating these pseudo-anomalies, ADP further optimizes the use of concepts with simple self-tuning scheme. ADP could outperform state-of-the-arts in recent few-shot benchmarks. Moreover, ADP can be applied in scenarios where text labels are scarce, without experiencing a significant drop compared to using the label. We believe our work could shed a light in exploring model personalization for downstream tasks beyond generative modeling.

References

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. GANomaly: Semi-supervised anomaly detection via ad-

versarial training. In *The 14th Asian Conference on Computer Vision*, pp. 622–637. Springer, 2018.

Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pp. 161–169. Springer, 2019.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4): 947–969, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Cohen, N. and Hoshen, Y. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489. Springer, 2021.

Deng, H. and Li, X. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Hengel, A. v. d. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

Gudovskiy, D., Ishizaka, S., and Kozuka, K. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 98–107, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratlin, M., and Wang, Y. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, 2022.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. *OpenCLIP*. Zenodo, July 2021. doi: 10.5281/zenodo.5143773. URL https://doi.org/10.5281/zenodo.5143773.

Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., and Dabeer, O. WinCLIP: Zero-/few-shot anomaly classification and segmentation. *arXiv preprint arXiv:2303.14814*, 2023.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664–9674, 2021.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Ristea, N.-C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., and Shah, M. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Rudolph, M., Wandt, B., and Rosenhahn, B. Same same but DifferNet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1907–1916, 2021.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4393–4402, 2018.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., and Rabiee, H. R. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14902–14912, 2021.

Sheynin, S., Benaim, S., and Wolf, L. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8495–8504, 2021.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Wu, J.-C., Chen, D.-J., Fuh, C.-S., and Liu, T.-L. Learning unsupervised Metaformer for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4369–4378, 2021.

Yang, M., Wu, P., Liu, J., and Feng, H. MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities. *arXiv preprint arXiv:2205.00908*, 2022.

Yi, J. and Yoon, S. Patch SVDD: Patch-level SVDD for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., and Wu, L. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

Zavrtanik, V., Kristan, M., and Skočaj, D. DRAEM – A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Zou, Y., Jeong, J., Pemula, L., Zhang, D., and Dabeer, O. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *Proceedings of the European Conference on Computer Vision*, 2022.