This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MultiADS: Defect-aware Supervision for Multi-type Anomaly Detection and Segmentation in Zero-Shot Learning

Ylli Sadikaj[*] [1 5 §], Hongkuan Zhou[* 2 3], Lavdim Halilaj[2],
Stefan Schmid[2], Steffen Staab[3 4], Claudia Plant[1 6]

[1]Faculty of Computer Science, University of Vienna, Vienna, Austria
[2]Bosch Corporate Research, Robert Bosch GmbH, Renningen, Germany
[3]University of Stuttgart, Stuttgart, Germany, [4]University of Southampton, Southampton, UK
[5]UniVie Doctoral School Computer Science, University of Vienna, Vienna, Austria
[6]ds:UniVie, Vienna, Austria

{ylli.sadikaj, claudia.plant}@univie.ac.at, steffen.staab@ki.uni-stuttgart.de
{hongkuan.zhou, lavdim.halilaj, stefan.schmid5}@de.bosch.com

Abstract

*Precise optical inspection in industrial applications is crucial for minimizing scrap rates and reducing the associated costs. Besides merely detecting if a product is anomalous or not, it is crucial to know the distinct types of defects, such as a bent, cut, or scratch. The ability to recognize the "exact" defect type enables automated treatments of the anomalies in modern production lines. Current methods are limited to solely detecting whether a product is defective or not, without providing any insights into the defect type, but nevertheless detecting and identifying multiple defects. We propose MultiADS, a zero-shot learning approach, able to perform **Multi**-type **A**nomaly **D**etection and **S**egmentation. The architecture of MultiADS comprises CLIP and extra linear layers to align the visual and textual representation in a joint feature space. To the best of our knowledge, our proposal is the first approach to perform a multi-type anomaly segmentation task in zero-shot learning. Contrary to the other baselines, our approach i) generates specific anomaly masks for each distinct defect type, ii) learns to distinguish defect types, and iii) simultaneously identifies multiple defect types present in an anomalous product. Additionally, our approach outperforms zero/few-shot learning SoTA methods on image-level and pixel-level anomaly detection and segmentation tasks on five commonly used datasets: MVTec-AD, Visa, MPDD, MAD, and Real-IAD.*

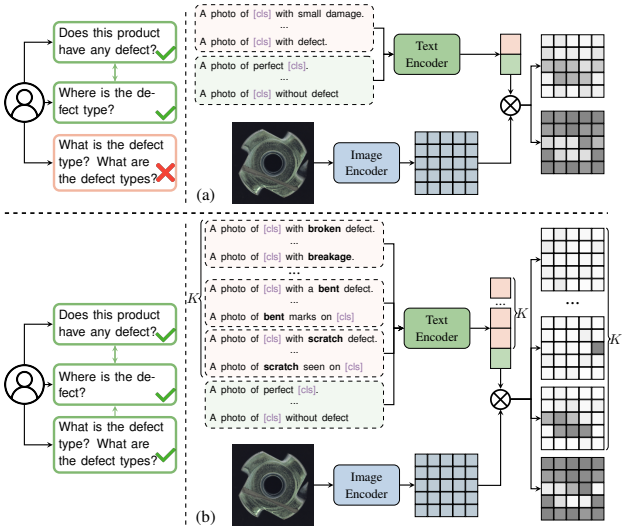
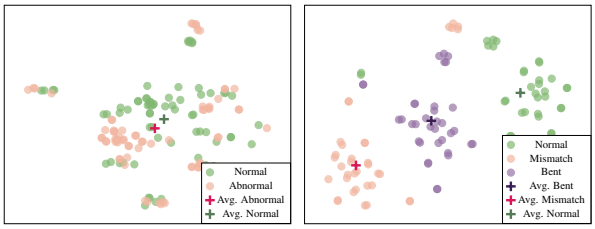


Figure 1. Comparison of common approaches and our approach: a) Common approaches typically differentiate only between normal and abnormal states; whereas b) our approach identifies $K+1$ states: one normal state and K distinct abnormal states corresponding to different defect types. This allows our method to distinguish between various defect types.

1. Introduction

One of the primary objectives of the manufacturing industries is to utilize their assembly lines for a wide range

[*]Both authors contributed equally to this work.
[§]Work done during PhD Sabbatical at Bosch Corporate Research.



(a) Common Approaches (b) Our Approach

Figure 2. Visualization of text prompts (TP) embeddings of common approaches and ours for Bracket Brown product of the MPDD dataset utilizing visualization tool t-SNE [34]. Dot signs $(\cdot)$ represent TP embeddings, plus signs $(+)$ represent the average embedding of TPs with the same color.

of product types. Modern factories are equipped with sophisticated and adaptable mechanisms allowing for a quick reconfiguration to various scenarios [19]. By doing so, the probability of outputting defective products is significantly increased. Therefore, to achieve intelligent manufacturing and prevent downtimes, rework, or quality losses, it is essential to detect anomalies promptly and with high precision [17, 30]. More concretely, identifying the specific defect* type in a product helps operators to understand the underlying causes and effectively implement preventive measures. In this regard, optical inspection via visual anomaly detection and segmentation is crucial to identify abnormal products and locate anomalous regions.

Recent approaches utilize prior knowledge in pre-trained models like CLIP [26] or DINO [4] to boost the generalization performance across a wide range of products for anomaly detection. CLIP-based approaches, such as [5, 15, 42], employ CLIP knowledge and adapt it for anomaly detection and segmentation by defining text-prompts for normal and abnormal states (cf. Figure 1a). Next, they compare the similarity between the image embedding and the average text embedding from generic sets of good and bad prompts. Thus, they are not exploiting anomaly-relevant knowledge, such as defect types, embedded in pre-trained vision language models (VLMs). On the other hand, fine-tuning in the specific domain often leads to overfitting on the training dataset [38], causing the model to lose valuable knowledge critical for accurate anomaly detection and segmentation. In Figure 2a, we visualize how averaging normal and abnormal text embeddings can lead to significant information loss.

In this paper, we present MultiADS, a zero-shot learning approach for multi-type anomaly detection and segmentation that leverages the prior knowledge of the common defect types in VLMs. It aligns the image embedding and the mean text embedding from a general set of good prompts and defect-specific sets of bad prompts. As illustrated in Figure 1b, through our approach, we can answer correctly all three questions, including the question regarding the defect type. Figure 2b shows that MultiADS preserves the meaningful semantic representation within the latent space and clearly distinguishes normal state and distinct defect types. Contrarily, competitive baselines could fail to separate between normal and abnormal states, as shown in Figure 2a. We conduct experiments on five datasets for anomaly detection and anomaly classification, MVTec [1], VisA [44], MPDD [16], MAD (real and simulated) [41], and Real-IAD [35]. We conducted evaluations in both zero-shot/few-shot settings. The empirical results demonstrate that incorporating defect-type information into the learning pipeline improves anomaly detection and segmentation performance across these five datasets. We summarize the key contributions as follows:

- Our MultiADS detects multiple defects of the same and/or different types in an anomalous product. Thus, we propose a new task, namely a multi-type anomaly detection and segmentation task, that aims to determine the defect type at the pixel level. We position MultiADS as a baseline in such a new task.
- We show that by leveraging anomaly-specific knowledge in pre-trained VLMs, MultiADS further improves its detection and segmentation performance.
- We present a Knowledge Base for Anomalies (KBA), that enhances the description of defect types. It can be utilized for defect-aware text prompt construction and facilitates the fine-tuning process of VLMs for anomaly detection and segmentation.
- Additionally, we evaluate the performance of MultiADS on anomaly detection and segmentation against 12 baselines both zero-shot/few-shot settings. The code implementation is publicly available at: https://github.com/boschresearch/MultiADS.

2. Related Work

In this section, we review the most relevant literature based on their learning paradigms and highlight how our approach distinguishes itself from existing methods.

Unsupervised Anomaly Detection. There exists a wide variation in the characteristics of objects and their defects, including differences in color, texture, size, and shape. This heterogeneity leads to an extensive range of defect types, making it challenging to compile a representative set of anomaly samples for training data. Thus, unsupervised anomaly detection approaches, such as [2, 13, 27, 37], require only normal images for train-

*We use *defect* and *anomaly* terms interchangeably.

ing. These methods typically model images without anomalies and classify any deviations from the learned representation as anomalies.

Zero-Shot Anomaly Detection (ZSAD). Recent studies have leveraged the power of large-scale VLMs such as CLIP [26] to perform anomaly detection without any target-specific training. The success of prompt learning in natural language processing has inspired methods such as CoOp [40] and CoCoOP [39], which automatically learn task-specific prompt contexts from only a few labeled examples. Early methods such as WinCLIP [15] and April-GAN [5] adapt CLIP by designing text prompts that differentiate "normal" from "abnormal" states. Also, they introduce window-based strategies or additional linear layers to enhance image segmentation performance.

Other approaches apply the same differentiation technique while adapting the construction for text prompt states. Thus, AnomalyCLIP [42] learns object-agnostic text prompts to capture generic cues of abnormality, SimCLIP [7] further adopts implicit prompt tuning. Similarly, FiLo [10] and AdaCLIP [3] enhance localization by replacing generic anomaly descriptions with adaptively learned fine-grained prompts or tuning hybrid learnable prompts by combining static and dynamic prompts. Contrary to other models, ClipSAM [20] proposes a novel collaboration between CLIP and SAM [18], whereas MuSc [22] detects anomalies by exploiting mutual scoring across unlabeled test images.

Few-Shot Anomaly Detection (FSAD). FSAD models, such as [12, 28, 29, 31], include several normal sample images from the target domain to train their model. PromptAD [23] refines the image–text alignment process by concatenating normal prompts with anomaly-specific suffixes. GraphCore [36] employs graph neural networks to capture rotation-invariant features from limited normal samples, while KAGprompt [32] constructs a kernel-aware hierarchical graph among multi-layer visual features. Other methods adopt reconstruction or feature-matching strategies—such as FastRecon [8] and FOCT [33]-to reconstruct normal appearances from a limited set of normal samples. Given the scarcity of anomalous samples, Anomalydiffusion [11] proposes to employ a latent diffusion model along with spatial anomaly embeddings to generate authentic anomaly image–mask pairs. Meanwhile, AnomalyGPT [9] is an interactive method integrating VLMs to provide defect-specific descriptions for a context-aware inspection. AnomalyDINO [6] uses DINOv2 [25] to extract robust patch-level features for FSAD.

A major limitation of existing vision-language ZSAD and FSAD methods is their binary focus—only distinguishing between normal and abnormal states, as illustrated in Figures 1 and 2. In contrast, MultiADS is designed to perform multi-type anomaly segmentation by constructing defect-specific text prompts that capture rich semantic attributes. This allows MultiADS to not only detect whether an image is anomalous but also to segment and classify the specific type of defect present - a capability that is critical for automated optical inspection in industrial applications.

3. Preliminaries

Here, we introduce the preliminary definitions of binary and multi-type anomaly detection and segmentation, as well as the backbone model.

3.1. Binary Detection and Segmentation

Let $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{target}}$ denote two different datasets, training and target datasets, respectively. Both datasets consist of X, Y, where $X = \{\mathbf{x}_i\}_{i=1}^{N}$ with N images, and $Y = \{(\mathbf{M}_i, y_i)\}_{i=1}^{N}$ with ground truth labels. Each image $\mathbf{x}_i \in \mathbb{R}^{H\times W}$ is masked with $\mathbf{M}_i$ and labeled with y_i, where $y_i \in \{0, 1\}$ is the indicator for anomaly or not and $\mathbf{M}_i \in \{0, 1\}^{H\times W}$ represents the binary anomaly map. Binary anomaly detection and segmentation (BADS) aim to determine if the given image $\mathbf{x}$ contains anomalies and also locate regions in an image that contain anomalies.

3.2. Multi-type Anomaly Segmentation

$\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{target}}$ denote the training and target datasets, respectively. Both datasets consist of X, Y', where $X = \{\mathbf{x}_i\}_{i=1}^{N}$ with N images and $Y' = \{\mathbf{M}'_i\}_{i=1}^{N}$. Each image $\mathbf{x}_i$ is labeled with $\mathbf{M}'_i \in \{0, 1, .., K\}^{H\times W}$, representing the multi-defect segmentation map for one normal class and K abnormal classes. Multi-type anomaly segmentation (MTAS) aims to locate the anomalies and identify various anomaly types.

3.3. Backbone Model

Contrastive Language Image Pre-training (CLIP) is a large-scale vision-language model pre-trained on million-scale image-text pairs, $\{(x_i, t_i)\}_{i=1}^{N}$. It encompasses an image feature encoder, $f(\cdot)$, and a text feature encoder, $g(\cdot)$. CLIP aims to maximize the correlation between $f(x_i)$ and $g(t_i)$ utilizing cosine similarity. Thus, for a given image input x and a closed set of text $T = \{t_1, \ldots, t_K\}$, representing the text prompt for K classes, CLIP performs classification as follows:

$$p(y = j|x) := \frac{exp(\langle f(x), g(t_j)\rangle/\tau)}{\sum_{i=1}^{K} exp(\langle f(x), g(t_i)\rangle/\tau)}, \quad (1)$$

where $\tau > 0$ is the temperature hyperparameter, whereas $\langle \cdot, \cdot \rangle$ represents the cosine similarity.

4. MultiADS Approach

Our proposed approach is a CLIP-based model adapted for zero-shot and few-shot learning for detecting anomalies and identifying the defect types in images from the manufacturing domain. It learns the alignment of image features with their corresponding text features that represent a distinct defect type, as shown in Figures 1 and 3. Anomaly maps constructed for each distinct defect type enable multi-class defect detection and segmentation.

Knowledge Base for Anomalies. We leverage the meta-data from established industrial defect detection datasets, including MVTec-AD, VisA, MPDD, MAD (real and simulated), and Real-IAD, to acquire comprehensive defect-aware information for each product class. Additionally, we incorporate supplementary defect-type properties (attributes) into our knowledge base for anomalies (KBA), including size and shape.

Initially, we group the defect types into superclasses, such that *bent*, *bent lead*, and *bent wire* are represented by the *bent* superclass, similarly *scratch*, *scratch head*, and *scratch neck* are under *scratch*. Thus, we have abstract classes like *bent*, *cut*, *scratch*, capturing all possible defect types that can occur in a given dataset. Details of the acquired information for all datasets part of our KBA are given in the Appendix.

Defect-aware Text Prompts. Next, we utilize the constructed KBA as prior knowledge for our text-prompt construction, as illustrated in Figures 1b and Figure 3. We select the same set of variations of text samples as in [5, 15] to construct text prompts for each given defect class. Figure 2 shows the difference between other baselines and our approach regarding the text prompt embeddings. More details for defect-aware text prompts are provided in the Appendix.

4.1. Training Phase

An overview of the training phase of our proposed method is shown in Figure 3 (LHS). We use different datasets for training and testing with their respective prompt set numbers denoted by K_1 and K_2.

4.1.1. Image and Text Embedding

Each image $\mathbf{x}$ is provided as input to the image encoder to get image patch embeddings at m different stages during encoding, as in [5, 42], $\mathbf{E}_i^p \in \mathbb{R}^{h \times w \times N_i}, i \in \{0, 1, ..., m\}$ with the resolution $h \times w$ and layer N_i, as well as one global image embedding $\mathbf{z}^x \in \mathbb{R}^{N_z}$. We use $K_1 + 1$ sets of text prompts: one representing the normal state and K_1 representing abnormal states corresponding to K_1 defect types. Each set of text prompts is fed into the CLIP text encoder, and we obtain an averaged text embedding for each set by averaging the embeddings of individual prompts. This process yields $K_1 + 1$ averaged text embeddings $\mathbf{z}^t \in \mathbb{R}^{N_z}$, each representing a distinct state.

4.1.2. Aligning Image Patches and Text Prompts

The visual encoder of CLIP is originally trained to align the global object embeddings with text embeddings. To align the two embedding spaces, visual - extracted by the CLIP image encoder, and textual - extracted by the CLIP text encoder, we utilize adapters consisting of a single linear learnable layer. For image patch embeddings at each stage i, a linear adapter takes $\mathbf{E}_i^p$ as input and outputs $\mathbf{Z}_i^p \in \mathbb{R}^{h \times w \times N_z}$. They are compared with $K_1 + 1$ text embeddings $\mathbf{z}^t$ to get the similarity map. Since we choose image patches embeddings at m different stages, we get m similarity maps $\mathbf{S}_i \in \mathbb{R}^{(K_1+1) \times h \times w}$, where h, w are the resolution of the similarity maps, K_1 is the number of defect types. Each map $\mathbf{S}_i$ is up-sampled to match the size of the input image and aligned with the ground truth segmentation map $\mathbf{M}'_x$.

4.1.3. Training Objective

Two typical losses, focal [24] and dice [21], are used for segmentation tasks. Focal loss is designed to address class imbalance issues, especially in tasks like object detection, where there is often a significant imbalance between classes. We face the same challenge, i.e., a high number of normal images and a low number of abnormal images; therefore, we apply a multi-class focal loss for multi-defect segmentation along with the binary dice loss for anomaly segmentation. These two training objectives are combined to form the final loss function:

$$\mathcal{L} = \sum_{i=1}^{m} \mathcal{L}_{\text{focal}}(UP(\mathbf{S}_i), \mathbf{M}'_x) + \mathcal{L}_{\text{dice}}(\mathbf{1} - UP(\mathbf{S}_i)[0], \mathbf{M}_x), \quad (2)$$

where $\mathbf{M}'_x$ represents the ground truth multi-defect segmentation map, and $\mathbf{M}_x$ is the binary anomaly map. $UP(\cdot)$ denotes the up-sampling function used to scale the similarity map to the input image resolution. Note that in the training phase, the global anomaly score a_x is not fine-tuned.

4.2. Inference Phase

To test the trained model's performance in the target dataset, we first construct $K_2 + 1$ sets of text prompts, representing one normal state without defect and K_2

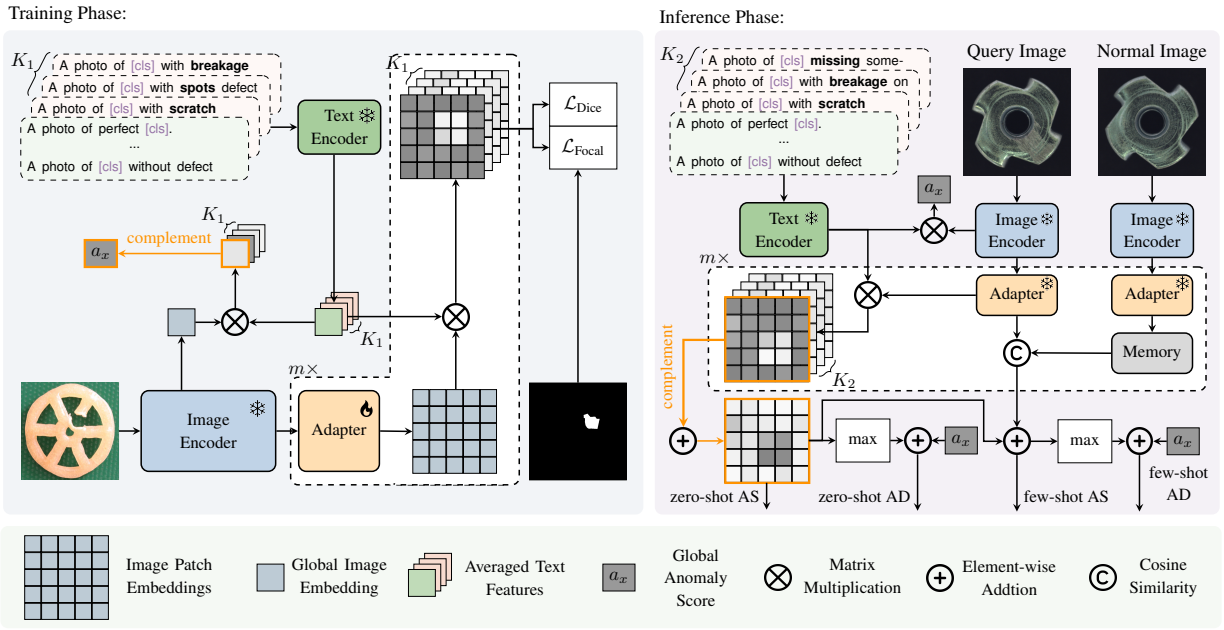


Figure 3. Training phase: K_1 text prompts describing the defect types plus one for good products are encoded into K_1+1 averaged text embeddings. The image patches are encoded and compared to these embeddings to produce K_1+1 similarity maps. For multi-type anomaly segmentation, we use dice and focal loss. Inference phase: we construct K_2+1 sets of text prompts. For anomaly segmentation (AS), we up-sample the complement of the normal layer's similarity map. For anomaly detection (AD), the global anomaly score a_x and the maximum score from the anomaly map are utilized. In few-shot testing, the query image is then compared with multiple reference (normal) images in the testing dataset to generate a similarity map. This similarity map is finally up-sampled and combined with the anomaly map for segmentation and classification tasks.

states representing distinct defect types of the target domain. An overview of the inference phase of our proposed method is shown in Figure 3 (RHS).

Each set of text prompts is input into the CLIP text encoder to generate embeddings, while the query image is passed through the CLIP image encoder and then the adapter to produce m similarity maps $\mathbf{S}_i \in \mathbb{R}^{(K_2+1)\times h\times w}$. The respective similarity maps are then up-sampled to match the original size of the input image. The multi-defect segmentation map is calculated by averaging the up-sampled similarity map:

$$\hat{\mathbf{M}}'_x = \frac{1}{m}\sum_{i=1}^{m} UP(\mathbf{S}_i). \tag{3}$$

We only take the *first* layer of similarity maps and perform a complement operation on each pixel to create the anomaly score map. Since there are m similarity maps, we average the m anomaly score maps to obtain the final anomaly map:

$$\hat{\mathbf{M}}_x = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1} - UP(\mathbf{S}_i)[0]. \tag{4}$$

The global image embedding $\mathbf{z}^x$ from the pre-trained CLIP image encoder is also compared with K_2+1 text embeddings to get K_2+1 global similarity scores. After the normalization, the complement of the similarity score compared to the normal state text prompts is used as the final global anomaly score a_x. We perform zero-shot learning based on the acquired anomaly map $\hat{\mathbf{M}}_x$ and global anomaly score a_x. Few-shot learning is conducted based on the acquired anomaly map $\hat{\mathbf{M}}_x$, global anomaly score a_x, and reference anomaly map $\hat{\mathbf{M}}_{\text{ref}}$ between query image and reference normal image(s).

4.2.1. Multi-type Anomaly Segmentation

The m similarity maps $\boldsymbol{S}_i, i \in \{1,\ldots,m\}$, are up-sampled to match the input image size and then averaged to produce the multi-defect segmentation map, $\hat{\mathbf{M}}'_x \in \mathbb{R}^{(K_2+1)\times h\times w}$. This map captures both the anomaly locations and their respective defect types, enabling effective support for the multi-type anomaly segmentation task.

4.2.2. Zero-shot Learning

For zero-shot learning, the output anomaly map $\hat{\mathbf{M}}_x$ is used for anomaly segmentation and compared with the ground truth labels. The highest anomaly score: $\max(\hat{\mathbf{M}}_x)$ on anomaly map and global anomaly score

a_x are averaged and then compared against a threshold θ to determine whether the image contains an anomaly.

4.2.3. Few-shot Learning

To conduct few-shot learning, we need to compute an extra reference anomaly map based on the similarity between the query image and several reference normal images. The reference normal image(s) are fed into the image encoder to get m stages of image patch embeddings. We leverage memory banks [5] to store the features of the reference images, which can be compared with input image features by cosine similarity to obtain the reference anomaly map $\hat{M}_{\text{ref}}$. The final anomaly map $\hat{\mathbf{M}}_{\text{final}} = \frac{1}{2}(\hat{\mathbf{M}}_x + \hat{\mathbf{M}}_{\text{ref}})$ is used for anomaly segmentation. $\hat{\mathbf{M}}_{\text{final}}$ instead of $\hat{\mathbf{M}}_x$ is used to determine the anomaly itself.

4.2.4. Filtering Out Product-irrelevant Defect Types

For a specific product type, only certain defect types are relevant. During the inference phase, this filtering step involves excluding text prompt sets associated with defect types that are not applicable to the product, ensuring that only relevant defect types are considered. Here, the method that includes this filtering process is referred to as MultiADS-F, while the original version without filtering remains as MultiADS.

5. Experiments

In this section, we describe datasets and baselines and discuss the results of the conducted experiments.

5.1. Datasets

Five common datasets: MVTec-AD [1], VisA [44], MPDD [16], MAD (simulated and real) [41], and Real-IAD [35] are used for the multi-type anomaly segmentation as well as the binary anomaly detection and segmentation task, respectively. More details of these datasets are provided in the Appendix.

5.2. Experiment Setting

We adopt a transfer learning setting, where the model is trained on one of the datasets and evaluated on the remaining. In the zero-shot learning scenario, the trained model is directly applied to the target dataset without any additional information from the target dataset. In contrast, the few-shot learning scenario allows the trained model to access a small number of normal images from the target dataset for further adaptation.

We use the ViT-L-14-336 CLIP backbone from OpenCLIP [14], pre-trained on the LAION-400M_E32 setting of open-clip. The learning rate is set to 0.001, with a batch size of 8. The stage number $m = 4$. The features are selected from layers: 6, 12, 18, and 24.

5.3. Evaluation Metrics

We assess the anomaly detection performance on zero/few-shot learning settings with three metrics, namely the receiver-operator curve (AUROC), the F1-score at the optimal threshold (F1-max), and the average precision (AP). Similar to [5, 15, 42], the anomaly segmentation is quantified by AUROC, F1-max, and the per-region overlap (PRO) of the segmentation using the pixel-wise anomaly scores. For the multi-type anomaly segmentation task, we employ AUROC, F1-score, and AP with the macro averaging setting.

5.4. Baselines

We compare the performance of our approach with the following 12 baselines: CLIP [26], CLIP-AC [26], CoOp [40], CoCoOp [39], PatchCore [28], WinCLIP [15], April-GAN [5], InCTRL [43], PromptAD [23], AnomalyCLIP [42], AdaCLIP [3], and AnomalyGPT [9]. CLIP, CLIP-AC, CoCo, CoCoOP, WinCLIP, April-GAN, AnomalyCLIP, and AdaCLIP are zero-shot learning approaches. Whereas CoOp, WinCLIP, and April-GAN can also learn in the few-shot setting, as other approaches, PatchCore, PromptAD, InCTRL, and AnomalyGPT. The comparison of batch zero-shot setting with MuSc [22] and AnomalyDINO [6] is discussed in the Appendix. We did not include other baselines such as [7, 10, 20] because their authors did not provide implementation yet.

In the evaluation process, we use the basic approach, MultiADS, and the filtering-based variant, MultiADS-F.

5.5. Results

Next, we present and discuss results from the experiments for multi-type anomaly segmentation in zero-shot settings and binary ZSAD and FSAD.

5.5.1. Multi-type Anomaly Segmentation

First, we discuss our MultiADS's performance in the new task, the multi-type anomaly segmentation (MTAS) task, which can segment various defect types. To the best of our knowledge, we are the first to perform such a task, and thus we present MultiADS as a baseline.

Table 1 shows the results of MultiADS on the MTAS task in a zero-shot learning setting. We observe that our approach achieves high accuracy in terms of the AUROC metric for pixel-level segmentation of distinct defects in all datasets. As expected, MultiADS performs with higher accuracy in terms of AP metric on datasets

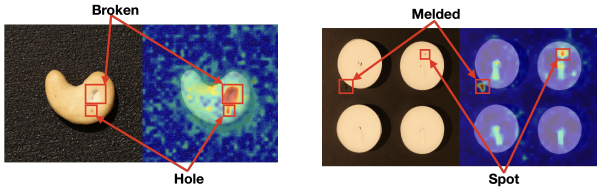
Table 1. Results on MTAS Task of MultiADS.

Train	Target	Pixel-Level		
		AUROC	F1-score	AP
MVTec-AD	VisA	93.6	22.3	24.8
	MPDD	95.2	42.8	53
	MAD-sim	92.1	27.9	31.5
	MAD-real	89.2	52.5	52.3
	Real-IAD	89.5	22.6	25.0
VisA	MVTec-AD	89.1	24	30.5
	MPDD	95.3	46.7	50.5
MPDD	VisA	93.4	22.1	23.3
	MVTec-AD	89.4	23.9	27.6
Real-IAD	MVTec-AD	87.7	21.4	29.9
	VisA	88.1	23.8	24.8

with fewer anomaly types, such as MPDD and MAD-real, and the accuracy is slightly lower on datasets with multiple anomaly types appearing concurrently, such as Real-IAD and VisA. Additionally, we found that MultiADS performs slightly better on the VisA dataset when our model is trained on the MVTec-AD or Real-IAD datasets rather than the MPDD dataset due to higher similarity between defect types of the VisA dataset with MVTec-AD and Real-IAD datasets. Similarly, the VisA dataset serves as a good model trainer regarding the performance of the model on the MVTec-AD dataset. In summary, these results indicate that MultiADS can successfully differentiate between various defect types. We provide more results on the MTAS task in the Appendix.

Multi-type Anomaly Awareness. Figure 4 shows that multiple defect types, such as *broken* and *hole*, can appear on one image, and MultiADS can successfully locate and classify these defects. Additionally, in Table 2, we listed the segmentation performance for some sample defect types that are seen/unseen during the training phase. We notice that defects such as *holes* and *damages* are relatively easy to locate and classify because they also occur on the training dataset - MVTec-AD. It may be that these defects are similar in terms of shape to those they have in datasets. For unseen defects like *extra* and *stuck*, our model achieves slightly lower accuracy. On the other hand, for other unseen defects such as *pit*, we can still perform with high accuracy on the classification task. These results reflect that our approach has generalization ability even on large and complex datasets and unseen defects in the training dataset.

Ablation Study. We present the results of our ablation studies on MTAS, quantifying the contributions of the KBA component. As Table 3 shows, the performance improves with the detailed text prompts constructed by KBA in both VisA and MAD-sim datasets. Similar patterns are present across all datasets.



(a) Broken and Hole defects. (b) Melded and Spot defects.

Figure 4. MultiADS locates and identifies simultaneously multi-type anomalies on cashew (a) and candle (b) products.

Table 2. Results MTAS for zero-shot setting at pixel-level for sample defect-types. The model is trained on the MVTec-AD dataset. - indicates **unseen** defect types while ✓indicates **seen** defect types during training.

(a) VisA

	Defects	AUROC	F1-Score	AP
-	Extra	94.07	2.11	0.15
-	Stuck	91.54	10.51	7.76
✓	Bent	96.53	6.07	7.74
✓	Hole	99.55	12.64	25.19

(b) Real-IAD

	Defects	AUROC	F1-Score	AP
-	Pit	97.08	6.15	1.01
✓	Contamin.	90.03	6.12	1.86
✓	Scratch	92.63	4.37	2.96
✓	Damage	96.61	6.31	9.75

Table 3. Ablation studies on the role of KBA for MTAS

	MVTec → VisA			MVTec → MAD-sim		
KBA	AUROC	F1-score	AP	AUROC	F1-score	AP
-	87.0	22.1	23.6	91.1	25.1	26.5
✓	93.6	22.3	24.8	92.1	27.9	31.5

5.5.2. Binary Detection and Segmentation

ZSAD. In Table 4, we show the performance on ZSAD for pixel-level (AUROC, AUPRO) and image-level (AUROC, AP) on VisA, MPDD, MAD (sim and real), and Real-IAD datasets. We selected these metrics to evaluate the performance following [42]. For a fair comparison, our approach and baseline approaches, including WinCLIP, April-GAN, AnomalyCLIP, and AdaCLIP, are trained on the MVTec-AD dataset. We observe that MultiADS and MultiADS-F are the best overall performers, especially when performance is evaluated with the AUPRO and AUROC metrics at the pixel and image levels, respectively. We note that our approach achieves the best performance for all metrics on both levels on the recent datasets, MAD and Real-IAD, which are even more challenging. Meanwhile, MultiADS-F is the best overall performer on the MPDD, MAD-real, and Real-IAD datasets, indicating that text prompts of non-relevant defect types present more noise for these datasets. Note that MultiADS and MultiADS-F have the same scores for the MAD-sim dataset, as all defect types appear for all product types. The best baseline performer is the AnomalyCLIP approach.

Table 5 shows the ablation study quantifying the contributions of KBA, global anomaly score, and different stage numbers on the ZASD task. The stage number has

Table 4. Zero-shot anomaly detection and segmentation. (Bold represents best performer; underline indicates second best performer, * means results are taken from papers)

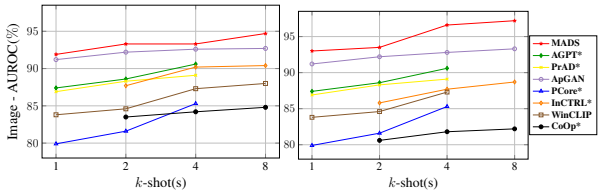
ZSAD			Pixel-Level		Image-Level	
Dataset	Method	Venue	AUROC	AUPRO	AUROC	AP
VisA	CLIP*	ICML21	46.6	14.8	66.4	71.5
	CLIP-AC*	ICML21	47.8	17.3	65.0	70.1
	CoOp*	IJCV22	24.2	3.8	62.8	68.1
	CoCoOp*	CVPR22	93.6	-	78.1	-
	WinCLIP	CVPR23	79.6	56.8	78.1	81.2
	April-GAN	CVPR23	94.2	86.8	78.0	81.4
	AnomalyCLIP	CVPR24	**95.5**	87.0	82.1	85.4
	AdaCLIP	ECCV24	95	-	75.4	79.3
	MultiADS (ours)		95	**89.7**	**83.6**	**86.9**
	MultiADS-F (ours)		94.5	87.4	82.5	86.5
MPDD	CLIP*	ICML21	62.1	33.0	54.3	65.4
	CLIP-AC*	ICML21	58.7	29.1	56.2	66.0
	CoOp*	IJCV22	15.4	2.3	55.1	64.2
	CoCoOp*	CVPR22	95.2	-	61	-
	WinCLIP	CVPR23	76.4	48.9	63.6	69.9
	April-GAN	CVPR23	94.1	83.2	73.0	80.2
	AnomalyCLIP	CVPR24	**96.5**	88.7	77.0	**82.0**
	AdaCLIP	ECCV24	96.3		66.3	75
	MultiADS (ours)		95.8	**89.7**	78.3	78.4
	MultiADS-F (ours)		96.3	89.5	**79.7**	80.5
MAD-sim	WinCLIP	CVPR23	77.6	55.8	54.3	90.2
	April-GAN	CVPR23	80.4	61.5	56	91
	AnomalyCLIP	CVPR24	77.9	40.1	54.6	90.9
	AdaCLIP	ECCV24	85.7	-	55.2	90.5
	MultiADS (ours) MultiADS-F (ours)		**88.0**	**74.2**	**57.1**	**94.4**
MAD-real	WinCLIP	CVPR23	60.5	26.9	64.1	87.6
	April-GAN	CVPR23	88.2	69.5	62.9	87.7
	AnomalyCLIP	CVPR24	88.3	65.1	66.8	90
	AdaCLIP	ECCV24	85.7	-	59	86.5
	MultiADS (ours)		89.7	74.0	78.3	**92.9**
	MultiADS-F (ours)		**90.7**	**75.2**	**78.5**	**92.9**
Real-IAD	WinCLIP	CVPR23	87.1	59.9	75	72.3
	April-GAN	CVPR23	96	86.8	75.7	73.5
	AnomalyCLIP	CVPR24	96.2	85.7	78.4	76.7
	AdaCLIP	ECCV24	95.3	-	70.1	68.5
	MultiADS (ours)		**96.6**	87.1	**78.7**	**79.1**
	MultiADS-F (ours)		96.3	**87.2**	78.2	78.5

the highest impact; the drop in performance is around 5% in terms of AP for both datasets when $m = 3$.

Table 5. Ablation studies on the role of KBA, global anomaly score a_x, and stage number m on the ZSAD task. Pixel-level results are ignored since a_x is only used at the image-level.

ZSAD			MVTec → VisA				MVTec → MPDD			
			Pixel-Level		Image-Level		Pixel-Level		Image-Level	
m	a_x	KBA	AUROC	AUPRO	AUROC	AP	AUROC	AUPRO	AUROC	AP
3	✓	✓	94.5	87.7	79.5	82.3	93.7	84.3	68.2	74.8
4	-	✓	-	-	82.1	85.8	-	-	76.5	78.1
4	✓	-	94.4	88.7	82.4	86.1	95.7	89.1	77.9	77.6
4	✓	✓	95.0	89.7	83.6	86.9	95.8	89.5	78.3	78.4

FSAD. Figure 5 shows the results for the FSAD task, for image-level (AUROC) with different numbers of shots, $k = [1, 2, 4, 8]$, on the Visa and MVTec-AD datasets. Similarly to ZSAD, we train our model on the MVTec-AD dataset and test on VisA and vice versa. We note that the most competitive baselines are April-GAN, PromptAD, and AnomalyGPT. We observe that MultiADS is the best overall performer for both datasets. The same performance patterns are found on other datasets,



(a) VisA dataset (b) MVTec-AD dataset

Figure 5. Few-Shot Image level (AUROC) accuracy for different k-shots on the VisA and MVTec-AD datasets. (* - results taken from papers, AGPT - AnomalyGPT, PCore - PatchCore, PrAD - PromptAD, ApGAN - April-GAN)

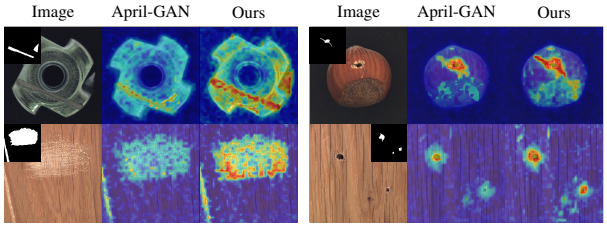


Figure 6. Visualization of anomaly segmentation from VisA and MVTec-AD datasets, in the few-shot (k=4) for defect types - scratch and hole. Each anomaly is highlighted to illustrate the ability of April-GAN and MultiADS.

too. The main advantage of our approach lies in extending the investigation based on defect awareness, supporting our claim that the main drawback of other methods is the two-state (normal and abnormal) limitation.

Figure 6 depicts a qualitative evaluation of the FSAD results of MultiADS and the best overall competitor, April-GAN, for scratch and hole defect types. We observe that MultiADS demonstrates higher confidence in identifying anomalies and achieves better segmentation across the same and different defect types due to its enhanced ability to capture the semantics of different defect types. More results are provided in the Appendix.

6. Conclusion

In this paper, we propose MultiADS, which constructs defect-aware text prompts to improve the performance of anomaly detection and segmentation tasks. We present a multi-type anomaly segmentation task that aims to determine the defect types and locations at the pixel level. We evaluated MultiADS on such a new task and positioned it as a baseline that can be used by the community. Finally, we evaluate MultiADS's performance against 12 baselines in ZSAD/FSAD on five datasets. Our evaluation demonstrates that MultiADS achieves the best performance in most cases for ZSAD/FSAD. In the future, we plan to explore adapting our approach to learn text prompt embeddings.

7. Acknowledgement

This work was partially funded by the European Union's Horizon RIA research and innovation programme under grant agreement No. 101092908 (SMARTEDGE). The authors also thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Hongkuan Zhou.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019. 2, 6

[2] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022. 2

[3] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, 2024. 3, 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640. IEEE, 2021. 2

[5] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 2, 3, 4, 6

[6] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. *CoRR*, abs/2405.14529, 2024. 3, 6

[7] Chenghao Deng, Haote Xu, Xiaolu Chen, Haodi Xu, Xiaotong Tu, Xinghao Ding, and Yue Huang. Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 1761–1770, New York, NY, USA, 2024. Association for Computing Machinery. 3, 6

[8] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17435–17444, 2023. 3

[9] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. *arXiv preprint arXiv:2308.15366*, 2023. 3, 6

[10] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024. 3, 6

[11] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8526–8534, 2024. 3

[12] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, page 303–319, Berlin, Heidelberg, 2022. Springer-Verlag. 3

[13] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *Computer Vision – ECCV 2022*, pages 303–319, Cham, 2022. Springer Nature Switzerland. 2

[14] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 6

[15] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, pages 19606–19616. IEEE, 2023. 2, 3, 4, 6

[16] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 66–71, 2021. 2, 6

[17] Yu Jiang, Wei Wang, and Chunhui Zhao. A machine vision-based realtime anomaly detection method for industrial products using deep learning. In *2019 Chinese Automation Congress (CAC)*, pages 4842–4847, 2019. 2

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[19] Tomas Kliestik, Marek Nagy, and Katarina Valaskova. Global value chains and industry 4.0 in the context of lean workplaces for enhancing company performance and its comprehension via the digital readiness and expertise of workforce in the v4 nations. *Mathematics*, 11 (3), 2023. 2

[20] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam

collaboration for zero-shot anomaly segmentation. *Neurocomputing*, 618:129122, 2025. 3, 6

[21] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 465–476. Association for Computational Linguistics, 2020. 4

[22] Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *International Conference on Learning Representations*, 2024. 3, 6

[23] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16838–16848, 2024. 3, 6

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 4

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 3, 6

[27] 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 (CVPR)*, 2022. 2

[28] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14298–14308, 2022. 3, 6

[29] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1906–1915, 2021. 3

[30] Patrick Ruediger-Flore, Matthias Klar, Marco Hussong, Avik Mukherjee, Moritz Glatt, and Jan C. Aurich. Comparing binary classification and autoencoders for vision-based anomaly detection in material flow. *Procedia CIRP*, 121:138–143, 2024. 2

[31] 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 (ICCV)*, pages 8495–8504, 2021. 3

[32] Fenfang Tao, Guo-Sen Xie, Fang Zhao, and Xiangbo Shu. Kernel-aware graph prompt learning for few-shot anomaly detection, 2025. 3

[33] Long Tian, Hongyi Zhao, Ruiying Lu, Rongrong Wang, Yujie Wu, Liming Wang, Xiongpeng He, and Xiyang Liu. Foct: Few-shot industrial anomaly detection with foreground-aware online conditional transport. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 6241–6249, New York, NY, USA, 2024. Association for Computing Machinery. 3

[34] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9:2579–2605, 2008. 2

[35] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. 2, 6

[36] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[37] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore, 2023. 2

[38] Hongkuan Zhou, Lavdim Halilaj, Sebastian Monka, Stefan Schmid, Yuqicheng Zhu, Bo Xiong, and Steffen Staab. Visual representation learning guided by multimodal prior knowledge. *CoRR*, abs/2410.15981, 2024. 2

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022. 3, 6

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 3, 6

[41] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad: A dataset and benchmark for pose-agnostic anomaly detection. *arXiv preprint arXiv:2310.07716*, 2023. 2, 6

[42] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt

learning for zero-shot anomaly detection. In *ICLR*. OpenReview.net, 2024. 2, 3, 4, 6, 7

[43] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17826–17836, 2024. 6

[44] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. *arXiv preprint arXiv:2207.14315*, 2022. 2, 6