AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection

Qihang Zhou[1*], **Guansong Pang**[2*], **Yu Tian**[3], **Shibo He**[1†], **Jiming Chen**[1†]
[1]Zhejiang University [2] Singapore Management University [3]Harvard University
[1] {zqhang, s18he, cjm}@zju.edu.cn [2]gspang@smu.edu.sg
[3]ytian11@meei.harvard.edu

Abstract

Zero-shot anomaly detection (ZSAD) requires detection models trained using auxiliary data to detect anomalies without any training sample in a target dataset. It is a crucial task when training data is not accessible due to various concerns, *e.g.*, data privacy, yet it is challenging since the models need to generalize to anomalies across different domains where the appearance of foreground objects, abnormal regions, and background features, such as defects/tumors on different products/organs, can vary significantly. Recently large pre-trained vision-language models (VLMs), such as CLIP, have demonstrated strong zero-shot recognition ability in various vision tasks, including anomaly detection. However, their ZSAD performance is weak since the VLMs focus more on modeling the class semantics of the foreground objects rather than the abnormality/normality in the images. In this paper we introduce a novel approach, namely AnomalyCLIP, to adapt CLIP for accurate ZSAD across different domains. The key insight of AnomalyCLIP is to learn object-agnostic text prompts that capture generic normality and abnormality in an image regardless of its foreground objects. This allows our model to focus on the abnormal image regions rather than the object semantics, enabling generalized normality and abnormality recognition on diverse types of objects. Large-scale experiments on 17 real-world anomaly detection datasets show that AnomalyCLIP achieves superior zero-shot performance of detecting and segmenting anomalies in datasets of highly diverse class semantics from various defect inspection and medical imaging domains. Code will be made available at `https://github.com/zqhang/AnomalyCLIP`.

1 Introduction

Anomaly detection (AD) has been widely applied in various applications, such as industrial defect inspection (Bergmann et al., 2019; Xie et al., 2023; Roth et al., 2022; Huang et al., 2022; Mou et al., 2022; Chen et al., 2022; Bergmann et al., 2020; Pang et al., 2021a; Reiss & Hoshen, 2023; You et al., 2022; Liznerski et al., 2020; Ding et al., 2022; Zhou et al., 2023; Cao et al., 2023) and medical image analysis (Pang et al., 2021a; Qin et al., 2022; Liu et al., 2023; Ding et al., 2022; Tian et al., 2021; 2023; Fernando et al., 2021). Existing AD approaches typically assume that training examples in a target application domain are available for learning the detection models (Pang et al., 2021b; Ruff et al., 2021). However, this assumption may not hold in various scenarios, such as i) when accessing training data violates data privacy policies (*e.g.*, to protect the sensitive information of patients), or ii) when the target domain does not have relevant training data (*e.g.*, inspecting defects in a manufacturing line of new products). Zero-shot anomaly detection (ZSAD) is an emerging task for AD in such scenarios, to which the aforementioned AD approaches are not viable, as it requires detection models to detect anomalies without any training sample in a target dataset. Since anomalies from different application scenarios typically have substantial variations in their visual appearance, foreground objects, and background features, *e.g.*, defects on the surface of one product vs. that on the other products, lesions/tumors on different organs, or industrial defects vs. tumors/lesions in medical images, detection models with strong generalization ability w.r.t.

[*]Equal contribution. † Corresponding authors.

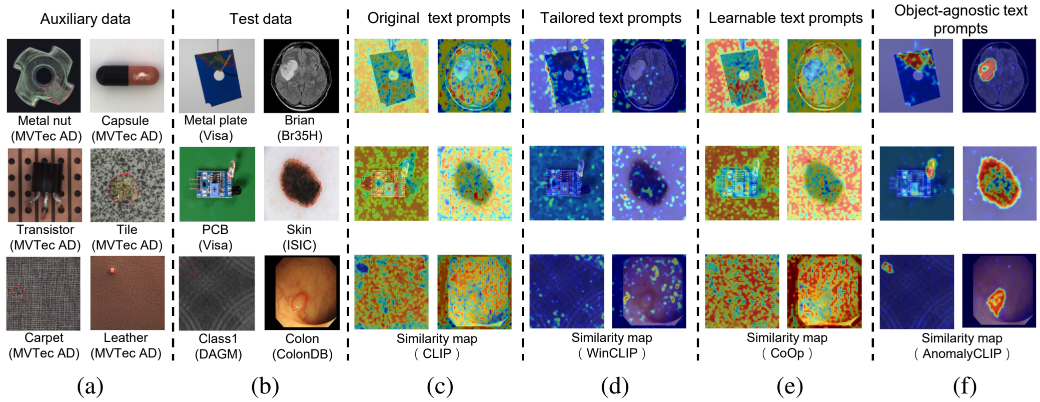

Figure 1: Comparison of ZSAD results on (b) test data using (c) original text prompts in CLIP (Radford et al., 2021), (d) tailored text prompts for AD in WinCLIP (Jeong et al., 2023), (e) learnable text prompts for general vision tasks in CoOp (Zhou et al., 2022a), and (f) object-agnostic text prompts in our AnomalyCLIP. (a) presents a set of auxiliary data we can use to learn the text prompts. The results are obtained by measuring the similarity between text prompt embeddings and image embeddings. The ground-truth anomaly regions are circled in red in (a) and (b). (c), (d), and (e) suffer from poor generalization across different domains, while our AnomalyCLIP in (f) can well generalize to anomalies in diverse types of objects from different domains.

such variations are needed for accurate ZSAD. Recently large pre-trained vision-language models (VLMs) (Radford et al., 2021; Kirillov et al., 2023) have demonstrated strong zero-shot recognition ability in various vision tasks, including anomaly detection (Jeong et al., 2023). Particularly, being pre-trained using millions/billions of image-text pairs, CLIP (Radford et al., 2021) has been applied to empower various downstream tasks (Zhou et al., 2022b; Rao et al., 2022; Khattak et al., 2023; Sain et al., 2023) with its strong generalization capability. WinCLIP (Jeong et al., 2023) is a seminal work in the ZSAD line, which designs a large number of artificial text prompts to exploit the CLIP's generalizability for ZSAD. However, the VLMs such as CLIP are primarily trained to align with the class semantics of foreground objects rather than the abnormality/normality in the images, and as a result, their generalization in understanding the visual abnormality/normality is restricted, leading to weak ZSAD performance. Further, the current prompting approaches, using either manually defined text prompts (Jeong et al., 2023) or learnable prompts (Sun et al., 2022; Zhou et al., 2022a), often result in prompt embeddings that opt for global features for effective object semantic alignment (Zhong et al., 2022; Wu et al., 2023), failing to capture the abnormality that often manifests in fine-grained, local features, as shown in Fig. 1d and Fig. 1e. In this paper we introduce a novel approach, namely AnomalyCLIP, to adapt CLIP for accurate ZSAD across different domains. AnomalyCLIP aims to learn object-agnostic text prompts that capture generic normality and abnormality in an image regardless of its foreground objects. It first devises a simple yet universally-effective learnable prompt template for the two general classes – normality and abnormality – and then utilizes both image-level and pixel-level loss functions to learn the generic normality and abnormality globally and locally in our prompt embeddings using auxiliary data. This allows our model to focus on the abnormal image regions rather the object semantics, enabling remarkable zero-shot capability of recognizing the abnormality that has similar abnormal patterns to those in auxiliary data. As shown in Fig. 1a and Fig. 1b, the foreground object semantics can be completely different in the fine-tuning auxiliary data and target data, but the anomaly patterns remain similar, *e.g.*, scratches on metal nuts and plates, the misplacement of transistors and PCB, tumors/lesions on various organ surfaces, etc. Text prompt embeddings in CLIP fail to generalize across different domains, as illustrated in Fig. 1c, but object-agnostic prompt embeddings learned by AnomalyCLIP can effectively generalize to recognize the abnormality across different domain images in Fig. 1f.

In summary, this paper makes the following main contributions.

- We reveal for the first time that learning object-agnostic text prompts of normality and abnormality is a simple yet effective approach for accurate ZSAD. Compared to current text prompting approaches that are primarily designed for object semantic alignment (Jeong et al., 2023; Zhou et al., 2022b), our text prompt embeddings model semantics of generic abnormality and normality, allowing object-agnostic, generalized ZSAD performance.

- We then introduce a novel ZSAD approach, called AnomalyCLIP, in which we utilize an object-agnostic prompt template and a glocal abnormality loss function (i.e., a combination of global and local loss functions) to learn the generic abnormality and normality prompts using auxiliary data. In doing so, AnomalyCLIP largely simplifies the prompt design and can effectively apply to different domains without requiring any change on its learned two prompts, contrasting to existing methods like WinCLIP whose effectiveness relies heavily on extensive engineering on hundreds of manually defined prompts.
- Comprehensive experiments on 17 datasets from various industrial and medical domains demonstrate that AnomalyCLIP achieves superior ZSAD performance of detecting and segmenting anomalies in datasets of highly diverse class semantics from defect inspection and medical imaging domains.

2 Preliminary

CLIP consists of a text encoder and visual encoder denoted as $T(\cdot)$ and $F(\cdot)$, respectively. Both encoders are mainstream multi-layer networks such as ViT (Dosovitskiy et al., 2020; Vaswani et al., 2017). Using text prompts is a typical way to achieve the embeddings of different classes for zero-shot recognition. Particularly, a text prompt template $\mathbb{G}$ with the class name c can be passed through $T(\cdot)$ to obtain its corresponding textual embedding $g_c \in \mathbb{R}^D$. The text prompt template commonly used in CLIP looks like `A photo of a [cls]`, where `[cls]` represents the target class name. Then $F(\cdot)$ encodes an image x_i to derive visual representations, where the class token $f_i \in \mathbb{R}^D$ is treated as its visual embedding (global visual embedding), and patch tokens $f_i^m \in \mathbb{R}^{H \times W \times D}$ are referred to as local visual embeddings. CLIP performs zero-shot recognition by measuring the similarity between textual and visual embeddings. In specific, given a target class set $\mathcal{C}$ and an image x_i, CLIP predicts the probability of x_i belonging to c as follows:

$$p(y = c|x_i) = P(g_c, f_i) = \frac{exp(< g_c, f_i > /\tau)}{\sum_{c \in \mathcal{C}} exp(< g_c, f_i >)/\tau)}, \tag{1}$$

where τ is a temperature hyperparameter, and the operator $< \cdot, \cdot >$ represents the computation of cosine similarity. Unlike many vision tasks that involve many objects and use the name of the objects as the class name `[cls]`, we posit that performing ZSAD tasks using CLIP should be object-agnostic, so we propose to design two classes of text prompts (*i.e.*, normality and abnormality) and compute the possibility of these two classes according to Eq. 1. We denote the probability of being abnormal $P(g_a, f_i)$ as the anomaly score. The computation is extended from global visual embeddings to local visual embeddings to derive the corresponding segmentation maps $S_n \in \mathbb{R}^{H \times W}$ and $S_a \in \mathbb{R}^{H \times W}$, where each entry (j, k) are computed as $P(g_n, f_i^{m(j,k)})$ and $P(g_a, f_i^{m(j,k)})$.

3 AnomalyCLIP: object-agnostic prompt learning

3.1 Approach overview

In this paper, we propose AnomalyCLIP to adapt CLIP to ZSAD via object-agnostic prompt learning. As shown in Fig. 2, AnomalyCLIP first introduces object-agnostic text prompt templates, where we design two generic object-agnostic text prompt templates of g_n and g_a to learn generalized embedding for the normality and abnormality classes, respectively (see Sec. 3.2). To learn such generic text prompt templates, we introduce global and local context optimization to incorporate global and fine-grained anomaly semantics into object-agnostic textual embedding learning. In addition, textual prompt tuning and DPAM are used to support the learning in the textual and local visual spaces of CLIP. Finally, we could integrate the multiple intermediate layers to provide more local visual details [1]. During training, all modules are jointly optimized by the combination of global and local context optimization. During inference, we quantify the misalignment of textual and global/local visual embeddings to obtain the anomaly score and anomaly score map, respectively (see Sec. 3.3).

[1] Although we use only the top feature map of visual encoder in the code implementation, AnomalyCLIP is capable of incorporating multiple intermediate feature maps.

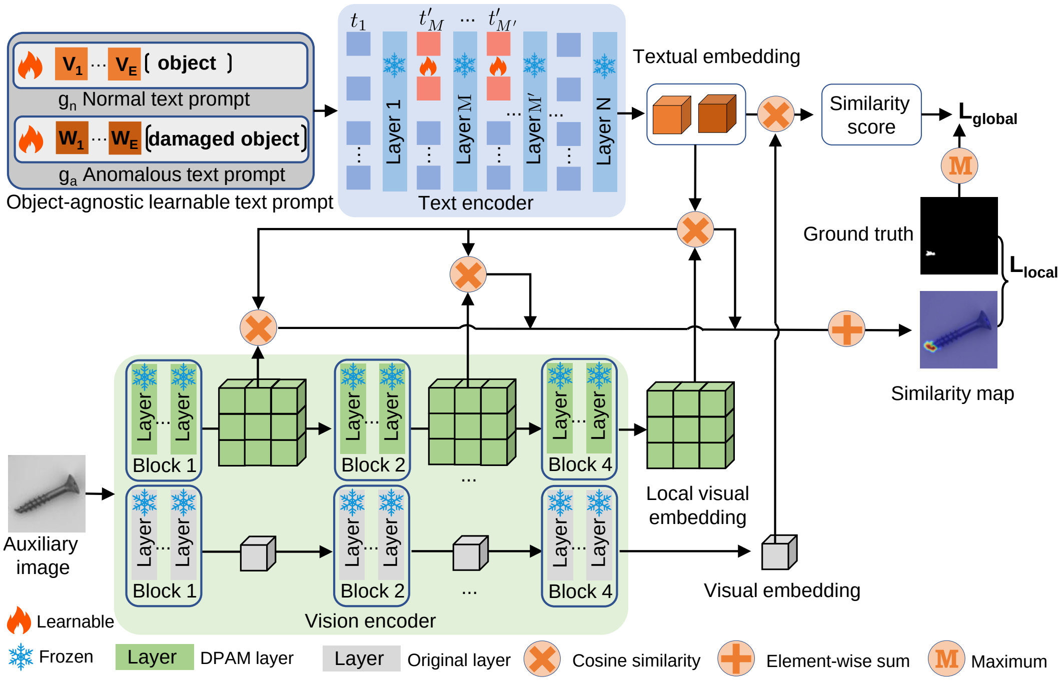

Figure 2: Overview of AnomalyCLIP. To adapt CLIP to ZSAD, AnomalyCLIP introduces object-agnostic text prompt templates to capture generic normality and abnormality regardless of the object semantics. Then, we introduce glocal context optimization to incorporate global and fine-grained anomaly semantics into object-agnostic text prompt learning. Finally, textual prompt tuning and DPAM are used to enable the prompt learning in the textual and local visual spaces of CLIP.

3.2 Object-agnostic text prompt design

Commonly used text prompt templates in CLIP, like `A photo of a [cls]`, primarily focus on object semantics. Consequently, they fail to generate textual embeddings that capture anomaly and normal semantics to query corresponding visual embeddings. To support the learning of anomaly-discriminative textual embeddings, we aim to incorporate prior anomaly semantics into text prompt templates. A trivial solution is to design the templates with specific anomaly types, such as `A photo of a [cls] with scratches` However, the pattern of anomaly is typically unknown and diverse, so it is practically difficult to list all possible anomaly types. Therefore, it is important to define text prompt templates with generic anomaly semantics. For this purpose, we can adopt the text `damaged [cls]` to cover comprehensive anomaly semantics, facilitating the detection of diverse defects such as scratches and holes. Nevertheless, utilizing such text prompt templates poses challenges in generating generic anomaly-discriminating textual embeddings. This is because CLIP's original pre-training focuses on aligning with object semantics instead of the abnormality and normality within images. To address this limitation, we can introduce learnable text prompt templates and tune the prompts using auxiliary AD-relevant data. During the fine-tuning process, these learnable templates can incorporate both broad and detailed anomaly semantics, resulting in textual embeddings that are more discriminative between normality and abnormality. This helps avoid the need for manually defined text prompt templates that require extensive engineering (Jeong et al., 2023). These text prompts are referred to as **object-aware text prompt templates** and defined as follows:

$$g_n = [V_1][V_2] \dots [V_E][cls]$$
$$g_a = [W_1][W_2] \dots [W_E][damaged][cls],$$

where $[V]_i$ and $[W]_i$ ($i \in 1, \dots, E$) are learnable word embeddings in normality and abnormality text prompt templates, respectively.

ZSAD tasks require models to detect anomalies in previously unseen target datasets. These datasets often exhibit significant variations in object semantics among different objects, like various defects on one product vs. another, or discrepancies between industrial defects and medical imaging tumors. However, despite these substantial differences in object semantics, the underlying anomaly patterns could be similar. For instance, anomalies like scratches on metal nuts and plates, or the misplacement of transistors and PCB, as well as tumors on the surface of various organs, can share similar anomaly patterns. We hypothesize that the key of accurate ZSAD is to identify these generic anomaly patterns regardless of the varying semantics of different objects. Therefore, the inclusion of object semantics in object-aware text prompt templates is often unnecessary for ZSAD. It can even hinder the detection of anomalies in classes that have not been seen during the learning process. More importantly, excluding the object semantics from text prompt templates allows learnable text prompt templates to focus on capturing the characteristics of anomalies themselves, rather than the objects. Motivated by this, we introduce object-agnostic prompt learning, with the aim to capture generic normality and abnormality within images regardless of the object semantics. Different from object-aware text prompt templates, as shown below, the **object-agnostic text prompt templates** replace the class name in g_n and g_a with `object`, blocking out the class semantics of objects:

$$g_n = [V_1][V_2]\dots[V_E][object]$$
$$g_a = [W_1][W_2]\dots[W_E][damaged][object].$$

This design empowers the object-agnostic text prompt template to learn the shared patterns of different anomalies. As a result, the generated textual embeddings are more generic and capable of identifying anomalies across diverse objects and different domains. Further, this prompt design is versatile and can be applied to different target domains without any modification, *e.g.*, requiring no knowledge about the object name or anomaly types in a target dataset.

3.3 LEARNING GENERIC ABNORMALITY AND NORMALITY PROMPTS

Glocal context optimization To effectively learn the object-agnostic text prompts, we devise a joint optimization approach that enables the normality and abnormality prompt learning from both global and local perspectives, namely global and local context optimization. The global context optimization aims to enforce that our object-agnostic textual embeddings are matched with the global visual embeddings of images of diverse objects. This helps effectively capture the normal/abnormal semantics from a global feature perspective. The local context optimization is introduced to enable object-agnostic text prompts to concentrate on fine-grained, local abnormal regions from M intermediate layers of the visual encoder, in addition to the global normal/abnormal features. Formally, let $\mathcal{M}$ be the set of intermediate layers used (*i.e.*, $M = |\mathcal{M}|$), our text prompts are learned by minimizing the following glocal loss function:

$$L_{total} = L_{global} + \lambda \textstyle\sum_{M_l \in \mathcal{M}} L_{local}^{M_l}, \tag{2}$$

where λ is a hyperparameter to balance the global and local losses. L_{global} is a cross-entropy loss that matches the cosine similarity between the object-agnostic textual embeddings and visual embeddings of normal/abnormal images from auxiliary data. Let $S \in \mathbb{R}^{H_{image} \times W_{image}}$ be the ground-truth segmentation mask, with $S_{jk} = 1$ if the pixel is as an anomaly and $S_{jk} = 0$ otherwise, then we have

$$S_{n,M_l}^{(j,k)} = P(g_n, f_{i,M_l}^{m(j,k)}),\ \ S_{a,M_l}^{(j,k)} = P(g_a, f_{i,M_l}^{m(j,k)}),\ \ \text{where}\ \ j \in [1, H], k \in [1, W]$$
$$L_{local} = Focal(Up([S_{n,M_l}, S_{a,M_l}]), S) + Dice(Up(S_{n,M_l}), I - S) + Dice(Up(S_{a,M_l}), S),$$

where $Focal(\cdot,\cdot)$ and $Dice(\cdot,\cdot)$ denote a focal loss (Lin et al., 2017) and a Dice loss (Li et al., 2019) respectively. The operators $Up(\cdot)$ and $[\cdot,\cdot]$ represent the unsampling and concatenation along with the channel, and I represents the full-one matrix. Since the anomalous regions are typically smaller than the normal ones, we use focal loss to address the imbalance problem. Furthermore, to ensure that the model establishes an accurate decision boundary, we employ the Dice loss to measure the overlaps between the predicted segmentation $Up(S_{n,M_l})$/$Up(S_{a,M_l})$ and the ground truth mask.

Refinement of the textual space To facilitate the learning of a more discriminative textual space via Eq. 2, inspired by Jia et al. (2022) and Khattak et al. (2023), we introduce randomly initialized learnable token embeddings in the text encoder to refine the textual space for its adaptation to AD. To control the degree of refinement in the textual space, we choose to replace the prefix part of the original token embeddings with learnable token embeddings in the text encoder, from the bottom (the second layer) to the top. In particular, we denote the token embeddings of the learnable text prompt as $t_m = \{t_m^1, t_m^2, \dots, t_m^P\}$, which we denote as $t_m = t_m^i|_1^P$

and m represents the layer index of the text encoder. We introduce additional multi-layer trainable tokens $q_m = q_m^i|_1^Q = \{q_m^1, q_m^2, \ldots, q_m^Q\}$, $Q < P$. To adapt the original textual representations of layer m, we replace $t_m^i|_1^Q$ with q_m, thus deriving the new token embeddings $t'_m = [q_m^i|_1^Q, t_m^i|_{Q+1}^P] = \{q_m^1, q_m^2, \ldots, q_m^Q, t_m^{Q+1}, \ldots, t_m^P\}$. Then, t'_m is forwarded into T_m to obtain $t_{m+1} = [r_{m+1}^i|_1^Q, t_{m+1}^i|_{Q+1}^P] = \{r_{m+1}^1, r_{m+1}^2, \ldots, r_{m+1}^Q, t_{m+1}^{Q+1}, \ldots, t_{m+1}^P\}$. To provide layer-wise adaptation, we discard the obtained $r_{m+1}^i|_1^Q$ and initialize new learnable token embeddings $q_{m+1}^i|_1^Q$. Through the concatenation operation, we obtain $t'_{m+1} = [q_{m+1}^i|_1^Q, t_{m+1}^i|_{Q+1}^P]$, which further refines the textual representations of layer $m+1$. We repeat this operation until we reach the designated layer M'. This procedure is given by:

$$
\begin{aligned}
t_2 &= T_1(t_1) \\
&\cdots \\
[r_{m+1}^i|_1^Q, t_{m+1}^i|_{Q+1}^P] &= T_m([q_m^i|_1^Q, t_m^i|_{Q+1}^P]) \\
[r_{m+2}^i|_1^Q, t_{m+2}^i|_{Q+1}^P] &= T_{m+1}([q_{m+1}^i|_1^Q, t_{m+1}^i|_{Q+1}^P]) \\
&\cdots \\
t_{M'+1} &= T_{M'}([r_{M'}^i|_1^Q, t_{M'}^i|_{Q+1}^P]),
\end{aligned}
\tag{3}
$$

where the operator $[\cdot, \cdot]$ represents the concatenation along the channel.

Refinement of the local visual space Since the visual encoder of CLIP is originally pre-trained to align global object semantics, the contrastive loss used in CLIP makes the visual encoder produce a representative global embedding for class recognition. Through the self-attention mechanism, the attention map in the visual encoder focuses on the specific tokens highlighted within the red rectangle in Fig 3b. Although these tokens may contribute to global object recognition, they disrupt the local visual semantics, which directly hinders the effective learning of the fine-grained abnormality in our object-agnostic text prompts (Li et al., 2023b). We found empirically that a diagonally prominent attention map helps reduce the disturbance from other tokens, leading to improved local visual semantics. Therefore, we propose a mechanism called Diagonally Prominent Attention Map to refine the local visual space, with the visual encoder kept frozen during training. To this end, we replace the original Q-K attention in the visual encoder with diagonally prominent attention, such as Q-Q, K-K, and V-V self-attention schemes. As demonstrated in Fig.3c, Fig.3d, and Fig. 3e, the refined DPAM attention maps are more diagonally prominent, resulting in substantially improved segmentation maps in both original CLIP and our AnomalyCLIP. Compared to CLIP that is based on global features and manually defined text prompts, the text prompts learned by AnomalyCLIP are more fine-grained, enabling substantially more accurate alignment between the normality/abnormality prompt embeddings and the local visual embeddings across four different self-attention schemes. This, in turn, allows AnomalyCLIP to generate accurate S_n and S_a for the joint optimization in Eq. 2. Unless otherwise specified, AnomalyCLIP utilizes V-V self-attention due to its superior overall performance. The performance of different self-attention mechanisms is analyzed in Sec. D. We also provide a detailed explanation about DPAM in Appendix C.

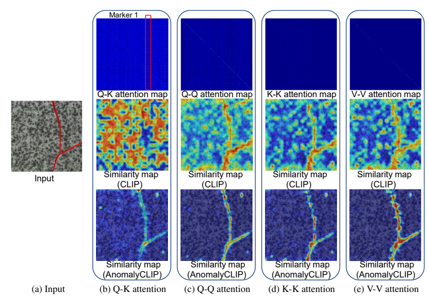

Figure 3: DPAM visualization.

Training and Inference During training, AnomalyCLIP minimizes the loss in Eq. 2 using an auxiliary AD-related dataset. As for inference, given a test image x_i, we use the similarity score $P(g_a, f_i)$ as the image-level anomaly score, with the anomaly score leaning toward one when the anomaly textual embedding g_a is aligned with global visual embedding f_i. For pixel-wise predictions, we merge the segmentation S_{n,M_l} and S_{a,M_l} of all selected intermediate layers, followed by an interpolation and smoothing operation. Formally, our anomaly score map $Map \in \mathbb{R}^{H_{image} \times W_{image}}$ is computed as $Map = G_\sigma(\sum_{M_l \in \mathcal{M}}(\frac{1}{2}(I - Up(S_{n,M_l})) + \frac{1}{2}Up(S_{a,M_l})))$, where G_σ represents a Gaussian filter, and σ controls smoothing.

Table 1: ZSAD performance comparison on industrial domain. The best performance is highlighted in red, and the second-best is highlighted in blue. † denotes results taken from original papers.

Task	Category	Datasets	$\|\mathcal{C}\|$	CLIP	CLIP-AC	WinCLIP	VAND	CoOp	AnomalyCLIP
Image-level (AUROC, AP)	Obj &texture	MVTec AD	15	(74.1, 87.6)	(71.5, 86.4)	(91.8, 96.5)†	(86.1, 93.5)†	(88.8, 94.8)	(91.5, 96.2)
	Obj	VisA	12	(66.4, 71.5)	(65.0, 70.1)	(78.1, 81.2)†	(78.0, 81.4)†	(62.8, 68.1)	(82.1, 85.4)
		MPDD	6	(54.3, 65.4)	(56.2, 66.0)	(63.6, 69.9)	(73.0, 80.2)	(55.1, 64.2)	(77.0, 82.0)
		BTAD	3	(34.5, 52.5)	(51.0, 62.1)	(68.2, 70.9)	(73.6, 68.6)	(66.8, 77.4)	(88.3, 87.3)
		SDD	1	(65.7, 45.2)	(65.2, 45.7)	(84.3, 77.4)	(79.8, 71.4)	(74.9, 65.1)	(84.7, 80.0)
	Texture	DAGM	10	(79.6, 59.0)	(82.5, 63.7)	(91.8, 79.5)	(94.4, 83.8)	(87.5, 74.6)	(97.5, 92.3)
		DTD-Synthetic	12	(71.6, 85.7)	(66.8, 83.2)	(93.2, 92.6)	(86.4, 95.0)	(-, -)	(93.5, 97.0)
Pixel-level (AUROC, PRO)	Obj &texture	MVTec AD	15	(38.4, 11.3)	(38.2, 11.6)	(85.1, 64.6)†	(87.6, 44.0)†	(33.3, 6.7)	(91.1, 81.4)
	Obj	VisA	12	(46.6, 14.8)	(47.8, 17.3)	(79.6, 56.8)†	(94.2, 86.8)†	(24.2, 3.8)	(95.5, 87.0)
		MPDD	6	(62.1, 33.0)	(58.7, 29.1)	(76.4, 48.9)	(94.1, 83.2)	(15.4, 2.3)	(96.5, 88.7)
		BTAD	3	(30.6, 4.4)	(32.8, 8.3)	(72.7, 27.3)	(60.8, 25.0)	(28.6, 3.8)	(94.2, 74.8)
		SDD	1	(39.0, 8.9)	(32.5, 5.8)	(68.8, 24.2)	(79.8, 65.1)	(28.9, 7.1)	(90.6, 67.8)
	Texture	DAGM	10	(28.2, 2.9)	(32.7, 4.8)	(87.6, 65.7)	(82.4, 66.2)	(17.5, 2.1)	(95.6, 91.0)
		DTD-Synthetic	12	(33.9, 12.5)	(23.7, 5.5)	(83.9, 57.8)	(95.3, 86.9)	(-, -)	(97.9, 92.3)

4 Experiments

4.1 Experiment setup

Datasets and Evaluation Metrics We conducted extensive experiments on 17 publicly available datasets, covering various industrial inspection scenarios and medical imaging domains (including photography, endoscopy, and radiology) to evaluate the performance of AnomalyCLIP. In industrial inspection, we consider MVTec AD (Bergmann et al., 2019), VisA (Zou et al., 2022), MPDD (Jezek et al., 2021), BTAD (Mishra et al., 2021), SDD (Tabernik et al., 2020), DAGM (Wieler & Hahn, 2007), and DTD-Synthetic (Aota et al., 2023). In medical imaging, we consider skin cancer detection dataset ISIC (Gutman et al., 2016), colon polyp detection datasets CVC-ClinicDB (Bernal et al., 2015), CVC-ColonDB (Tajbakhsh et al., 2015), Kvasir (Jha et al., 2020), and Endo (Hicks et al., 2021), thyroid nodule detection dataset TN3k (Gong et al., 2021), brain tumor detection datasets HeadCT (Salehi et al., 2021), BrainMRI (Salehi et al., 2021), Br35H (Hamada., 2020), and COVID-19 detection dataset COVID-19 (Chowdhury et al., 2020; Rahman et al., 2021). The SOTA competing methods include CLIP (Radford et al., 2021), CLIP-AC (Radford et al., 2021), WinCLIP (Jeong et al., 2023), VAND (Chen et al., 2023), and CoOp (Zhou et al., 2022b). We provide more details about the methods and data pre-processing in Appendix A. The anomaly detection performance is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC). Additionally, average precision (AP) for anomaly detection and AUPRO (Bergmann et al., 2020) for anomaly segmentation are also used to provide more in-depth analysis of the performance.

Implementation details We use the publicly available CLIP model[2] (`VIT-L/14@336px`) as our backbone. Model parameters of CLIP are all frozen. The length of learnable word embeddings E is set to 12. The learnable token embeddings are attached to the first 9 layers of the text encoder for refining the textual space, and λ is set to 4. We use the top feature map as local visual details for anomaly segmentation. Then, we fine-tune AnomalyCLIP using the test data on MVTec AD and evaluate the ZSAD performance on other datasets. As for MVTec AD, we fine-tune AomalyCLIP on the test data of VisA. We report dataset-level results, which are averaged across their respective sub-datasets. All experiments are conducted in PyTorch-2.0.0 with a single NVIDIA RTX 3090 24GB GPU. More details can be found in Appendix A.

4.2 Main results

ZSAD performance on diverse industrial inspection domains Table 1 shows the ZSAD results of AnomalyCLIP with five competing methods over seven industrial defect datasets of very different foreground objects, background, and/or anomaly types. AnomalyCLIP achieves superior ZSAD performance across the datasets, substantially outperforming the other five methods in most datasets. The weak performance of CLIP and CLIP-AC can be attributed to CLIP's original pre-training, which focuses on aligning object semantics rather than anomaly semantics. By using manually defined text prompts, WinCLIP and VAND achieve better results. Alternatively, CoOp adopts learnable prompts to learn the global anomaly semantics. However, those prompts focus on the global

[2]https://github.com/mlfoundations/open_clip

Table 2: ZSAD performance comparison on medical domain. The best performance is highlighted in red, and the second-best is highlighted in blue. Note that the image-level medical AD datasets do not contain segmentation ground truth, so the pixel-level medical AD datasets are different from the image-level datasets.

Task	Category	Datasets	$\|\mathcal{C}\|$	CLIP	CLIP-AC	WinCLIP	VAND	CoOp	AnomalyCLIP
Image-level (AUROC, AP)	Brain	HeadCT	1	(56.5, 58.4)	(60.0, 60.7)	(81.8, 80.2)	(89.1, 89.4)	(78.4, 78.8)	(93.4, 91.6)
		BrainMRI	1	(73.9, 81.7)	(80.6, 86.4)	(86.6, 91.5)	(89.3, 90.9)	(61.3, 44.9)	(90.3, 92.2)
		Br35H	1	(78.4, 78.8)	(82.7, 81.3)	(80.5, 82.2)	(93.1, 92.9)	(86.0, 87.5)	(94.6, 94.7)
	Chest	COVID-19	1	(73.7, 42.4)	(75.0, 45.9)	(66.4, 42.9)	(15.5, 8.5)	(25.3, 9.2)	(80.1, 58.7)
Pixel-level (AUROC, PRO)	Skin	ISIC	1	(33.1, 5.8)	(36.0, 7.7)	(83.3, 55.1)	(89.4, 77.2)	(51.7, 15.9)	(89.7, 78.4)
	Colon	CVC-ColonDB	1	(49.5, 15.8)	(49.5, 11.5)	(70.3,32.5)	(78.4, 64.6)	(40.5, 2.6)	(81.9, 71.3)
		CVC-ClinicDB	1	(47.5, 18.9)	(48.5, 12.6)	(51.2,13.8)	(80.5, 60.7)	(34.8, 2.4)	(82.9, 67.8)
		Kvasir	1	(44.6, 17.7)	(45.0, 16.8)	(69.7, 24.5)	(75.0, 36.2)	(44.1, 3.5)	(78.9, 45.6)
		Endo	1	(45.2, 15.9)	(46.6, 12.6)	(68.2, 28.3)	(81.9, 54.9)	(40.6, 3.9)	(84.1, 63.6)
	Thyroid	TN3K	1	(42.3, 7.3)	(35.6, 5.2)	(70.7, 39.8)	(73.6, 37.8)	(34.0, 9.5)	(81.5, 50.4)

feature and ignore the fine-grained local anomaly semantics, leading to their poor performance on anomaly segmentation. To adapt CLIP to ZSAD, AnomalyCLIP learns object-agnostic text prompts to focus on learning the generic abnormality/normality using global and local context optimization, enabling the modeling of both global and local abnormality/normality. Our resulting prompts can also generalize to different datasets from various domains. To provide more intuitive results, we visualize the anomaly segmentation results of AnomalyCLIP, VAND, and WinCLIP across different datasets in Fig. 4. Compared to VAND and WinCLIP, AnomalyCLIP can perform much more accurate segmentation for the defects from different industrial inspection domains.

Generalization from defect datasets to diverse medical domain datasets To evaluate the generalization ability of our model, we further examine the ZSAD performance of AnomalyCLIP on 10 medical image datasets of different organs across different imaging devices. Table 2 shows the results, where learning-based methods, including AnomalyCLIP, VAND and CoOp, are all tuned using MVTec AD data. It is remarkable that methods like AnomalyCLIP and VAND obtain promising ZSAD performance on various medical image datasets, even though they are tuned using a defect detection dataset. Among all these methods, AnomalyCLIP is the best performer due to its strong generalization brought by object-agnostic prompt learning. As illustrated in Fig. 4, AnomalyCLIP can accurately detect various types of anomalies in diverse medical images, such as skin cancer regions in photography images, colon polyps in endoscopy images, thyroid nodules in ultrasound images, and brain tumors in MRI images, having substantially better performance in locating the abnormal lesion/tumor regions than the other two methods WinCLIP and VAND. This again demonstrates the superior ZSAD performance of AnomalyCLIP in datasets of highly diverse object semantics from medical imaging domains.

Can we obtain better ZSAD performance if fine-tuned using medical image data? Comparing the promising performance in industrial datasets, AnomalyCLIP presents a relatively low performance in medical datasets. This is partly due to the impact of auxiliary data used in our prompt learning. So, then we examine whether the ZSAD performance on medical images can be improved if the prompt learning is trained on an auxiliary medical dataset. One challenge is that there are no available large 2D medical datasets that include both image-level and pixel-level annotations for our training. To address this issue, we create such a dataset based on ColonDB (More details see Appendix A), and then optimize the prompts in AnomalyCLIP and VAND using this dataset and evaluate their performance on the medical image datasets. The results are presented in Table 3. AnomalyCLIP and VAND largely improve their detection and segmentation performance compared to that fine-tuned on MVTec AD, especially for the colon polyp-related datasets such as CVC-ClincDB, Kvasir, and Endo (note that these datasets are all from different domains compared to the fine-tuning ColonDB dataset). AnomalyCLIP also exhibits performance improvement in detecting brain tumors in datasets such as HeadCT, BrainMRI, and Br35H. This is attributed to the visual similarities between colon polyps and brain tumors. Conversely, the symptom of the colon polyp differs significantly from that of diseased skin or chest, leading to performance degradation in ISIC and COVID-19. Overall, compared to VAND, AnomalyCLIP performs consistently better across all datasets of anomaly detection and segmentation.

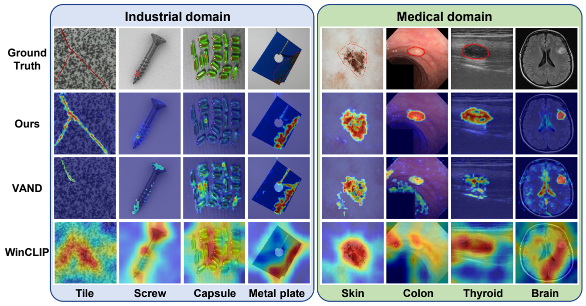

Figure 4: Segmentation visualization.

Table 3: ZSAD performance on medical images when fine-tuned by medical image datasets.

Category	Datasets	VAND	AnomalyCLIP
Classification			
	HeadCT	(89.1, 89.4)	(93.5, 95.1)
Brain	BrainMRI	(89.3, 90.9)	(95.5, 97.2)
	Br35H	(93.1, 92.9)	(97.9, 98.0)
Chest	COVID-19	(15.5, 8.5)	(70.9, 33.7)
Segmentation			
Skin	ISIC	(58.8, 31.2)	(83.0, 63.8)
	CVC-ClinicDB	(89.4, 82.3)	(92.4, 82.9)
Colon	Kvasir	(87.6, 39.3)	(92.5, 61.5)
	Endo	(88.5, 81.9)	(93.2, 84.8)
Thyroid	TN3K	(60.5, 16.8)	(79.2, 47.0)

Table 4: Module ablation.

Module	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
Base	(46.8, 15.4)	(66.3, 83.3)	(47.9, 17.1)	(54.4, 61.7)
$+T_1$	(68.4, 47.4)	(66.3, 83.3)	(54.8, 32.7)	(54.4, 61.7)
$+T_2$	(89.5, 81.2)	(90.8, 96.0)	(95.0, 85.3)	(81.7, 85.2)
$+T_3$	(91.1, 81.4)	(91.5, 96.2)	(95.5, 87.0)	(82.1, 85.4)

Table 5: Context optimization ablation.

Local.	Global.	MVTec AD		VisA	
		Pixel-level	Image-level	Pixel-level	Image-level
✗	✗	(71.7, 57.7)	(68.8, 85.8)	(74.7, 62.1)	(61.1, 69.1)
✗	✓	(80.3, 77.8)	(89.9, 95.4)	(86.6, 78.1)	(82.2, 84.9)
✓	✗	(91.0, 80.4)	(89.9, 96.0)	(95.2, 86.5)	(79.5, 83.2)
✓	✓	(91.1, 81.4)	(91.5, 96.2)	(95.5, 87.0)	(82.1, 85.4)

Object-agnostic vs. object-aware prompt learning To study the effectiveness of object-agnostic prompt learning in AnomalyCLIP, we compare AnomalyCLIP with its variant that uses an object-aware prompt template. The performance gain of AnomalyCLIP to its object-aware prompt learning variant is shown in Fig. 5, where positive values indicate our object-agnostic prompt templates are better than the object-aware one. It is clear that our object-agnostic prompt learning performs much better than, or on par with, the object-aware version in both image-level and pixel-level anomaly detection. This indicates that having object-agnostic prompts helps better learn the generic abnormality and normality in images, as the object semantics are often not helpful, or can even become noisy features, for the ZSAD task.

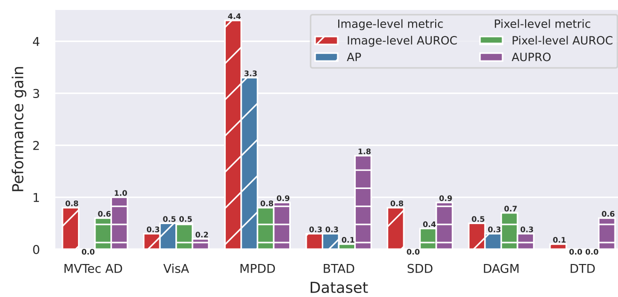

Figure 5: Performance gain of using object-agnostic prompts compared to object-aware prompts.

4.3 Ablation Study

Module ablation We first validate the effectiveness of different high-level modules of our AnomalyCLIP, including DPAM (T_1), object-agnostic text prompts (T_2), and added learnable tokens in text encoders (T_3). As shown in Table 4, each module contributes to the remarkable performance of AnomalyCLIP. DPAM improves the segmentation performance by enhancing local visual semantics (T_1). Object-agnostic text prompts focus on the abnormality/normality within images instead of the object semantics, allowing AnomalyCLIP to detect anomalies in diverse unseen objects. Therefore, introducing object-agnostic text prompts (T_2) significantly improves AnomalyCLIP. Furthermore, text prompt tuning (T_3) also brings performance improvement via the refinement of original textual space.

Context optimization Next we examine key modules in detail. The object-agnostic prompt learning is the most effective module, and it is driven by our glocal context optimization, so we consider two different optimization terms, local and global losses, in Eq. 2. The results are shown in Table 5. Both global and local context optimization contribute to the superiority of AnomalyCLIP. Global context optimization helps to capture global anomaly semantics, thus enabling more accurate image-level detection. Compared to global context optimization, local context optimization incorporates local anomaly semantics, which improves pixel-level performance and complements image-level performance. By synthesizing these two optimization strategies, AnomalyCLIP generally achieves better performance than using them individually.

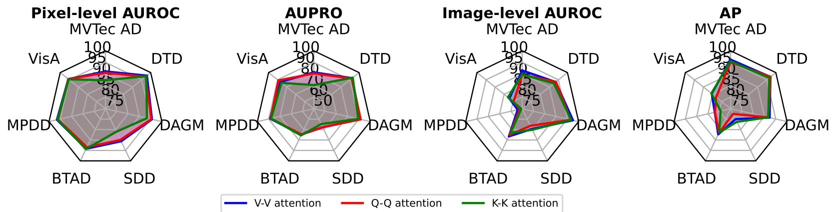

Figure 6: DPAM component ablation.

DPAM strategy ablation AnomalyCLIP uses V-V self-attention by default. Here we study the effectiveness of using two other DPAM strategies, including Q-Q and K-K self-attention, resulting in two AnomalyCLIP variants, namely AnomalyCLIP$_{qq}$ and AnomalyCLIP$_{kk}$. The comparison results are presented in Fig. 6. AnomalyCLIP$_{qq}$ achieves similar segmentation capabilities as AnomalyCLIP but suffers from degradation in detecting image-level anomalies. Conversely, while AnomalyCLIP$_{kk}$ performs well in anomaly classification, its segmentation performance is less effective than AnomalyCLIP and AnomalyCLIP$_{qq}$. The V-V self-attention is generally recommended in AnomalyCLIP. Detailed analysis of DPAM can be seen in Appendix C.

5 Related Work

Zero-shot anomaly detection ZSAD relies on the model's strong transferability to handle unseen anomalies (Aota et al., 2023). CLIP-AD (Liznerski et al., 2022) and ZOC (Esmaeilpour et al., 2022) are early studies in utilizing CLIP for ZSAD, but they mainly focus on the anomaly classification task. ACR (Li et al., 2023a) requires tuning on target-domain-relevant auxiliary data for ZSAD on different target datasets, while AnomalyCLIP can be applied to different datasets after it is trained on one general dataset. A very recent approach WinCLIP (Jeong et al., 2023) presents a seminal work that leverages CLIP for zero-shot classification and segmentation. It uses a large number of hand-crafted text prompts and involves multiple forward passes of image patches for anomaly segmentation. To tackle this inefficiency, VAND (Chen et al., 2023) introduces learnable linear projection techniques to enhance the modeling of local visual semantics. However, these approaches suffer from insufficiently generalized textual prompt embeddings, which degrades their performance in identifying anomalies associated with various unseen object semantics. AnomalyCLIP utilizes only two object-agnostic learnable text prompts to optimize the generic text prompts of abnormality and normality, and it can obtain segmentation results with just a single forward pass. AnomalyGPT (Gu et al., 2023) is a concurrent work in utilizing foundation models for AD, but it is designed for unsupervised/few-shot AD with manually crafted prompts.

Prompt learning Rather than resorting to full network fine-tuning, prompt learning emerges as a parameter-efficient alternative to achieve satisfactory results (Sun et al., 2022; Khattak et al., 2023; Kim et al., 2023; Zhou et al., 2022a). CoOp (Zhou et al., 2022b) introduces learnable text prompts for few-shot classification. On this basis, DenseCLIP (Rao et al., 2022) extends prompt learning to dense prediction tasks with an extra image decoder. Instead, AnomalyCLIP proposes object-agnostic prompt learning for anomaly detection, blocking out the potential adverse impact of the diverse object semantics on anomaly detection. Benefiting from the glocal context optimization, AnomalyCLIP can capture local anomaly semantics such that we can simultaneously perform classification and segmentation tasks without an additional decoder network like Rao et al. (2022).

6 Conclusion

In this paper, we tackle a challenging yet significant area of anomaly detection, ZSAD, in which there is no available data in the target dataset for training. We propose AnomalyCLIP to improve the weak generalization performance of CLIP for ZSAD. We introduce object-agnostic prompt learning to learn generic abnormality/normality text prompts for generalized ZSAD on image datasets of diverse foreground objects. Further, to incorporate global and local anomaly semantics into AnomalyCLIP, we devise a joint global and local context optimization to optimize the object-agnostic text prompts. Extensive experimental results on 17 public datasets demonstrate that AnomalyCLIP achieves superior ZSAD performance.

Acknowledgments

This work was supported by NSFC U1909207, NSFC 62088101 Autonomous Intelligent Unmanned Systems, and the Singapore Ministry of Education Academic Research Fund Tier 1 grant (21SISSMU031).

Reproducibility Statement

To ensure the reproducibility and completeness of this paper, we have included an Appendix consisting of five main sections. In Appendix A, we provide more implementation details of AnomalyCLIP, as well as the reproduction of other baseline methods. Appendix B provides key statistics about the datasets used in our experiments and the implementation of the auxiliary medical dataset for prompt tuning. Appendix D supplements the main paper with additional results and ablations. Further visualizations of similarity scores and maps are detailed in Appendix E. Additionally, the main paper presents only the average performance in each dataset that contains a number of data subsets, for which we present their fine-grained detection results, in Appendix F. Our code will be made publicly accessible once the paper is accepted.

References

Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5564–5572, 2023.

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*, pp. 9592–9600, 2019.

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*, pp. 4183–4192, 2020.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. *arXiv preprint arXiv:2303.13845*, 2023.

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.

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*, volume 36, pp. 383–392, 2022.

Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020. doi: 10.1109/ACCESS.2020.3010287.

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*, pp. 9737–9746, 2022.

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*, pp. 7388–7398, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 6568–6576, 2022.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7): 1–37, 2021.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 257–261. IEEE, 2021.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models, 2023.

David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), 2016.

A. Hamada. Br35h: Brain tumor detection 2020. *Online. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection*, 2020.

Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, pp. 263–274. Springer, 2021.

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*, pp. 303–319. Springer, 2022.

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*, pp. 19606–19616, 2023.

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)*, pp. 66–71. IEEE, 2021.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pp. 451–462. Springer, 2020.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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. *arXiv preprint arXiv:2304.02643*, 2023.

Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019.

Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023b.

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*, pp. 2980–2988, 2017.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. *arXiv preprint arXiv:2301.00785*, 2023.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *arXiv preprint arXiv:2205.11474*, 2022.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 01–06. IEEE, 2021.

Shancong Mou, Xiaoyi Gu, Meng Cao, Haoping Bai, Ping Huang, Jiulong Shan, and Jianjun Shi. Rgi: robust gan-inversion for mask-free image inpainting and unsupervised pixel-wise anomaly detection. In *The Eleventh International Conference on Learning Representations*, 2022.

Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021a.

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, 2021b.

Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022.

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*, pp. 8748–8763. PMLR, 2021.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18082–18091, 2022.

Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2155–2162, 2023.

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*, pp. 14318–14328, 2022.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2765–2775, 2023.

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*, pp. 14902–14912, 2021.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022.

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.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

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*, pp. 128–140. Springer, 2021.

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*, pp. 102930, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, volume 6, 2007.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15254–15264, 2023.

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.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.