Zero-Shot Anomaly Detection with Dual-Branch Prompt Selection

Zihan Wang[1,2]
zihan.wang5@mail.mcgill.ca
Samira Ebrahimi Kahou[2,3,4]
samira.ebrahimikahou@ucalgary.ca
Narges Armanfard [1,2]
narges.armanfard@mcgill.ca

[1] McGill University
Montreal, QC, Canada
[2] Mila – Quebec AI Institute
Montreal, QC, Canada
[3] University of Calgary
Calgary, AB, Canada
[4] CIFAR AI Chair
Canada

Abstract

Zero-shot anomaly detection (ZSAD) enables identifying and localizing defects in unseen categories by relying solely on generalizable features rather than requiring any labeled examples of anomalies. However, existing ZSAD methods, whether using fixed or learned prompts, struggle under domain shifts because their training data are derived from limited training domains and fail to generalize to new distributions. In this paper, we introduce PILOT, a framework designed to overcome these challenges through two key innovations: (1) a novel dual-branch prompt learning mechanism that dynamically integrates a pool of learnable prompts with structured semantic attributes, enabling the model to adaptively weight the most relevant anomaly cues for each input image; and (2) a label-free test-time adaptation strategy that updates the learnable prompt parameters using high-confidence pseudo-labels from unlabeled test data. Extensive experiments on 13 industrial and medical benchmarks demonstrate that PILOT achieves state-of-the-art performance in both anomaly detection and localization under domain shift.

1 Introduction

Anomaly detection and localization is critical in safety-critical applications, including industrial inspection and medical diagnostics, where timely identification of rare deviations from normal patterns can prevent costly failures and improve patient outcomes. Traditional anomaly detection methods include supervised approaches that require labeled anomalous examples and unsupervised techniques relying solely on normal data [6, 18, 23, 26, 27, 38]. Supervised methods require scarce and diverse anomaly labels, whereas unsupervised approaches avoid any need for anomalous samples by modeling only normal data, though they depend on comprehensive, representative normal data from the target domain [9]. Zero-shot Anomaly Detection (ZSAD) has emerged as a powerful approach by leveraging pretrained vision–language models (VLMs) such as CLIP [37], which learns joint representations of images and text from large-scale image–text data, to identify anomalies in visual data without requiring any target-domain annotations of normal or anomalous samples.

Early successful ZSAD approaches utilized CLIP by designing explicit textual prompts that described both the object category and its condition, such as "normal bottle" or "anomalous transistor". By crafting a large number of prompts that described relevant combinations of object category and condition, these methods enabled zero-shot classification and significantly outperformed vanilla CLIP baselines [11, 14, 20]. These approaches are known as fixed prompt methods because they rely on manually designed textual prompts. More recent methods have shifted towards a single learnable prompt, where the prompt itself is represented as a trainable embedding and optimized during training. This optimization is performed using auxiliary data, which refers to anomaly detection datasets that are separate from the target domain. For example, AnomalyCLIP [47] learns single object agnostic prompt embeddings to capture generic anomaly patterns, while AdaCLIP [8] further integrates visual knowledge into the prompt embeddings to enhance model adaptation.

Despite these advances, both fixed and single learnable prompt approaches remain vulnerable to domain shift, which refers to performance degradation when data distribution in the target domain differs significantly from training data. Although pretrained VLMs like CLIP have strong generalization ability, their performance can drop sharply when there are significant differences between the data seen during pretraining and the target environment [5, 16]. Even when learnable prompt embeddings are further trained on auxiliary anomaly detection datasets, that are different from the actual deployment domain, these models often struggle to represent the full diversity of possible anomalies or adapt to new categories encountered at test time. This limited generalization occurs for two main reasons: first, the auxiliary and target datasets may have significant distribution differences, as we empirically demonstrate in Appendix B.2; second, single prompt embeddings risk overfitting to the patterns and features specific to the auxiliary dataset, which further reduces their ability to handle unseen anomalies. As a result, prompts derived from narrowly defined auxiliary sources may fail to capture diverse visual features in the target data [33]. Therefore, to address these challenges, we introduce a framework that learns and dynamically weights multiple prompts, rather than relying on a fixed or single learnable prompt.

On the other hand, Test-Time Adaptation (TTA) techniques have demonstrated improved model robustness in various computer vision tasks by refining model parameters on unlabeled test data [29, 35, 39]. Although TTA is designed to address domain shift between training and test data, applying naive TTA strategies – without tailoring them to the specific task at hand – can inadvertently degrade performance. When used blindly, especially in tasks such as detection or localization, the adaptation process may optimize surrogate objectives that are misaligned with the true goals, often leading to suboptimal outcomes [30, 32, 41]. AnovL [13] was among the first to introduce TTA for anomaly localization refinement in ZSAD scenarios, but its focus is limited to localization and relies on fixed prompts. These limitations motivate us to explore how both anomaly detection and localization can be improved under domain shift by effectively utilizing unlabeled target data. [1].

To address these challenges, we propose a novel framework, **P**rompt Learning with **I**ntegrated **L**abel-Free Test-Time Adaptation for Zer**O**-Sho**T** Anomaly Detection (PILOT), a framework designed to mitigate the limitations of current ZSAD methods under domain shift. Instead of relying on a single learnable prompt, PILOT constructs a pool of learnable prompts, allowing the prompts to adaptively learn from various anomalies present in the aux-

[1] We present the Related Works section in Section A of the Supplementary Material.

iliary dataset. This design enables the model to adaptively weight the most relevant prompts for each input image, allowing for more nuanced representations of diverse anomalies and supporting more effective TTA. During inference, our label-free test-time adaptation strategy updates all learnable prompts according to their alignment with each test image, using high-confidence pseudo-labels generated from unlabeled target data. This targeted adaptation facilitates rapid and reliable adjustment to new domains (i.e., the test domain) without requiring any labeled data. In addition, to mitigate overfitting to the auxiliary dataset and enhance adaptation stability, PILOT incorporates an attribute memory bank, which consists of a curated set of fixed prompts that describe normal and anomalous states, leveraging pretrained semantic knowledge from the text encoder. This attribute memory bank serves as a semantic anchor, helping to stabilize localization performance during TTA. Together, these innovations enable PILOT to achieve robust anomaly detection and localization across a variety of challenging real-world scenarios. Our contributions are summarized as follows:

- We propose a dual-branch prompt learning framework that dynamically weight the most relevant prompts from both learnable prompt pool and attribute memory bank, enabling the model to leverage the most relevant anomaly cues for each input image.
- To our knowledge, this is the first label-free TTA strategy in ZSAD that leverages high-confidence pseudo-labels from unlabeled test samples to rapidly adapt prompt parameters at inference, yielding improvements in both anomaly detection and localization under domain shift.
- Extensive evaluation on industrial and medical benchmarks shows that our method outperforms existing approaches in both anomaly detection and localization metrics.

2 Proposed Method

ZSAD involves detecting and localizing anomalies in images from categories not encountered during the training phase. Formally, in the *training phase*, we employ an auxiliary dataset defined as $\mathcal{D}_{\text{aux}} = (I_i, y_i, G_i)_{i=1}^{|\mathcal{D}_{\text{aux}}|}$, where each image $I_i \in \mathbb{R}^{H \times W \times 3}$ is paired with an image-level label $y_i \in \{0,1\}$ (0 denotes normal and 1 denotes anomalous) and a pixel-level anomaly mask $G_i \in \{0,1\}^{H \times W}$. In the *inference phase*, we consider an unlabeled target dataset $\mathcal{D}_{\text{tar}} = \{I_t\}_{t=1}^{|\mathcal{D}_{\text{tar}}|}$, where each $I_t \in \mathbb{R}^{H \times W \times 3}$. For every I_t, the model outputs an anomaly score $\hat{y}_t \in [0,1]$, (with values closer to 1 indicating higher abnormality likelihood), along with a $\hat{G}_t \in [0,1]^{H \times W}$ that spatially localizes potential anomalous regions at pixel level. Crucially, the category sets of $\mathcal{D}_{\text{aux}}$ and $\mathcal{D}_{\text{tar}}$ are strictly disjoint, ensuring that the target domain remains unseen during training and thus establishing a true zero-shot setting.

To address the limitations of existing approaches, we propose a dual-branch framework that combines a learnable prompt pool $\mathcal{P}$ with an attribute memory bank $\mathcal{U}$. The training phase of our PILOT framework, depicted in Fig. 1, highlights these two key components. Further details on the inference phase are provided in Section 2.5, with in-depth descriptions of each component in the following sections and the complete pseudocode in Appendix E.

2.1 Learnable Prompt Pool

A successful recent approach for CLIP-based ZSAD employs a *single* learnable prompt, defined as a pair of sequences (p^n, p^a) representing normal and anomalous states, respectively,

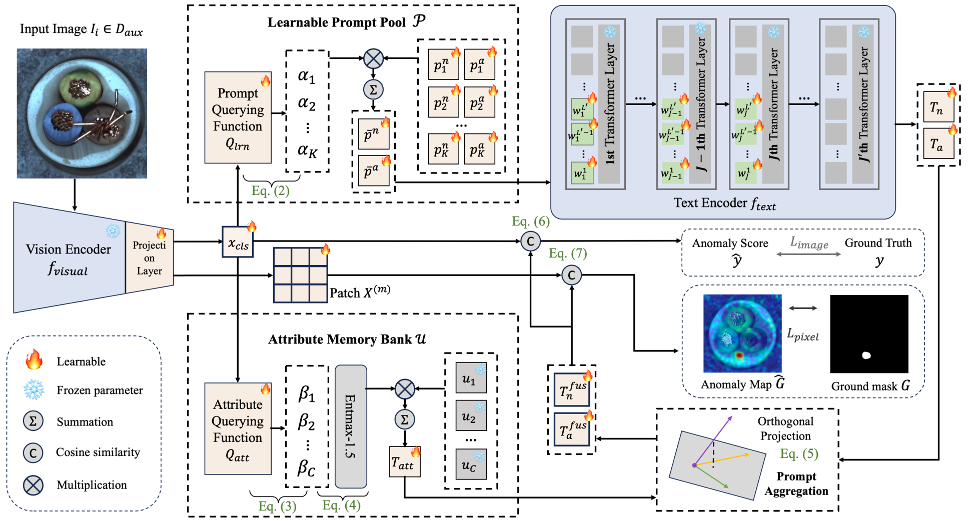

Figure 1: Overview of PILOT's training phase. The left panel shows the main workflow. The right panel details the Learnable Prompt Pool $\mathcal{P}$ and Attribute Memory Bank $\mathcal{U}$.

for binary anomaly classification [47]. Formally, this prompt can be represented as:

$$p^n = [v^1, \ldots, v^L, \text{"[normal state] object"}],\ p^a = [v^1, \ldots, v^L, \text{"[anomalous state] object"}] \quad (1)$$

where L is the number of learnable prefix vectors, and each $v^{(i)} \in \mathbb{R}^d$ is a d-dimensional vector trained jointly with the model. These vectors form a "soft prompt," intuitively acting as special tokens whose values steer the model's attention without modifying its core parameters, follows the prefix-tuning paradigm, which guides VLMs while keeping the main model weights frozen [24, 28]. The placeholders "[normal state]" and "[anomalous state]" filled with descriptors of the object's condition, such as "flawless" and "damaged."

Despite its simplicity and effectiveness, relying solely on a single learnable prompt for all input images limits generalization under domain shift, as discussed in the introduction, because a single pair of prompts often fails to adequately represent the diversity of anomaly patterns encountered in different scenarios. To address this limitation, we propose extending this single prompt to a learnable prompt *pool*, containing multiple learnable prompts, each associated with additional embeddings for adaptive weighting and alignment. Formally, the proposed prompt pool is defined as $\mathcal{P} = \{(p_k^n, p_k^a, s_k, r_k)\}_{k=1}^K$, where $K = |\mathcal{P}|$ is the cardinality of the prompt pool. Each (p_k^n, p_k^a) is constructed similarly to previous works, and the additional learnable embeddings s_k and r_k enable adaptive weighting of the prompts and alignment with the input image features. Here, s_k is a modulation vector for prompt k, applied element-wise to the image feature x_{cls} to adaptively emphasize information relevant to that prompt, while r_k provides a prompt-specific reference for measuring similarity. These embeddings facilitate an adaptive emphasis on prompts that best align with the diverse patterns present in the auxiliary dataset. To achieve this adaptive prompt weighting, we introduce a prompt-querying function, which computes prompt relevance scores given an image feature $x_{\text{cls}} \in \mathbb{R}^d$ produced by the CLIP vision encoder, i.e. f_{visual}, from an image I_i in $\mathcal{D}_{\text{aux}}$,:

$$Q_{\text{lrn}}(x_{\text{cls}}, \mathcal{P}) = \{\alpha_k := \text{sim}(x_{\text{cls}} \odot s_k, r_k)\}_{k=1}^K, \quad (2)$$

where "$\odot$" denotes element-wise multiplication, and $\mathrm{sim}(\cdot,\cdot)$ represents cosine similarity. Intuitively, α_k quantifies how well the image content aligns with the k-th prompt pair; this is realized by weighting the image feature x_{cls} with the attention mask s_k and measuring its similarity to the associated reference embedding r_k. The higher the score α_k is, the more similar the input image is to the k-th prompt (p_k^n, p_k^a). These scores are then used to form weighted combinations of the prompt token sequences as $\bar{p}^n = \sum_{k=1}^{K} \alpha_k p_k^n$, $\bar{p}^a = \sum_{k=1}^{K} \alpha_k p_k^a$. The resulting $\bar{p}^n$ and $\bar{p}^a$ are subsequently encoded using the CLIP pretrained text encoder to obtain the final text embeddings $T_n = f_{\mathrm{text}}(\bar{p}^n)$ and $T_a = f_{\mathrm{text}}(\bar{p}^a)$, which are employed for downstream anomaly scoring and localization. This process is visualized in Figure 1.

Although the text encoder f_{text} is frozen to preserve CLIP's rich pretrained textual representations, we employ a prefix-tuning strategy to enable effective fine-tuning and adaptation of CLIP. Inspired by similar prefix-tuning methods [24], we introduce additional learnable parameters to f_{text}. More specifically, as is shown in Figure 1, layer-wise prefix vectors are defined as $\left\{ w_j^{(\ell)} \in \mathbb{R}^d \mid j = 1, \dots, J,\ \ell = 1, \dots, L' \right\}$, where f_{text} consists of a total of J' transformer layers, and $J \leq J'$ denotes the number of initial layers to which learnable vectors are appended. L' represents the number of learnable vectors per layer. Specifically, these prefix vectors are prepended as additional tokens to the input sequence at the first J transformer layers, leaving the original transformer layers unchanged. At each layer j, these vectors $\{w_j^{(\ell)}\}_{\ell=1}^{L'}$ (illustrated as green blocks within the f_{text} module of Figure 1) are added before the standard, frozen token embeddings (represented as gray blocks in the the f_{text} module of Figure 1) prior to the self-attention mechanism in in each frozen transformer layer. This approach enables adaptation of the model's internal representations without modifying the pretrained CLIP parameters, thereby preserving their robustness and generalization capabilities. Furthermore, as is illustrated in Figure 1, PILOT fine-tunes only the final d-dimensional projection layer within the vision encoder f_{visual}. By restricting updates exclusively to this projection layer, we substantially reduce memory overhead and computational complexity. This targeted fine-tuning strategy also mitigates the risk of overfitting to auxiliary datasets, maintaining the strong generalization properties inherent to CLIP [15].

2.2 Attribute Memory Bank

Although the learnable prompt pool $\mathcal{P}$ aims to provide adaptive and diverse representations tailored to the auxiliary dataset, it may still be susceptible to overfitting when deployed on target domains. To mitigate this issue, we further introduce a fixed attribute memory bank comprising attribute embeddings extracted from the pretrained CLIP text encoder f_{text} (prior to applying our modifications). Recent works, such as WinCLIP [20], have demonstrated the robustness of fixed textual prompts by encoding multiple natural-language templates. Inspired by this, we construct our attribute memory bank from a set of descriptive natural-language templates (e.g., "A bright photo of a [State] object"), where *[State]* indicates descriptive attributes of the object's condition (e.g., "normal", "corroded", or "contaminated").

Each attribute template is passed through the frozen CLIP text encoder f_{text}, without any additional learnable tokens $w_j^{(\ell)}$, to produce a fixed attribute embedding, formally defined as $\mathcal{U} = \left\{ (u_c, g_c) \right\}_{c=1}^{C}$, where C is the number of attribute embeddings, and each $u_c \in \mathbb{R}^d$ represents the EOS-token embedding of the corresponding template. Unlike WinCLIP, which averages embeddings across multiple prompts, we associate each attribute embedding u_c with a learnable modulation vector g_c, similar to the learnable prompt pool. Similar to how s_k adapts the image features for each prompt, g_c dynamically modulates the contribution of each attribute according to the input image. Specifically, given an image feature x_{cls}, we introduce the attribute-querying function to compute attribute relevance scores as below:

$$Q_{\text{att}}\big(x_{\text{cls}}, \mathcal{U}\big) = \big\{\beta_c := \text{sim}\big(x_{\text{cls}} \odot g_c, u_c\big)\big\}_{c=1}^{C}, \tag{3}$$

where "$\odot$" denotes element-wise multiplication, and $\text{sim}(\cdot,\cdot)$ is the cosine similarity. Here, u_c itself serves as the reference embedding for each attribute, as it is obtained directly from the pretrained f_{text} and already contains rich semantic information. Each score β_c quantifies the alignment between the image and the c-th attribute embedding. The resulting scores $\{\beta_c\}_{c=1}^{C}$ quantify the alignment between the input and each attribute. Following [19], we inject i.i.d. Gumbel noise into the attribute logits and apply the entmax 1.5 operator [36] to obtain a sparse yet differentiable weighting over the most prominent cues:

$$\{\phi_c\}_{c=1}^{C} = \text{entmax}\,1.5(\beta_c + \text{noise}), \quad \sum_{c=1}^{C} \phi_c = 1, \quad \phi_c \geq 0. \tag{4}$$

We then compute the attribute text embedding as a weighted combination. $T_{att} = \sum_{c=1}^{C} \phi_c u_c$, $T_{att} \in \mathbb{R}^d$ The injected noise promotes exploration and prevents premature convergence to a single attribute, allowing the model to capture diverse and complementary semantic cues. By providing stable, pretrained semantic cues, the attribute memory bank acts as a semantic anchor during TTA discussed in Section 2.5. This mechanism helps to counteracts the potential overfitting and instability of learnable prompts, ensuring the model retains reliable semantic references even as it adapts to new domains.

2.3 Prompt Aggregation via Orthogonal Projection

Following [45], to robustly leverage both the stable semantics from the attribute embedding T_{att} and the adaptive semantics from the learnable prompts (T_n for normal and T_a for anomalous), we aggregate these representations using an orthogonal projection operation instead of simple averaging or concatenation. For each prompt embedding $T_{(\cdot)}$ (where $(\cdot) \in \{n, a\}$), we decompose it into components parallel and orthogonal to T_{att}:

$$T_{(\cdot)}^{\text{fus}} = T_{\text{att}} + (T_{(\cdot)} - \text{Proj}_{T_{\text{att}}}(T_{(\cdot)})), \quad \text{Proj}_{\eta}(\psi) = \frac{\eta^{\top}\psi}{\|\eta\|^2}\eta \tag{5}$$

This construction anchors the fused embedding in the robust attribute space, while enriching it with only those cues from the learnable prompt that are *complementary* to the attribute. Geometrically, $T_{(\cdot)}^{\text{fus}}$ preserves all the core semantics provided by T_{att}, and selectively incorporates novel information from $T_{(\cdot)}$ that lies in the orthogonal complement. Further analysis can be found in Appendix E.8. This parameter-free aggregation prevents redundancy and enhances robustness to distributional shift, especially during TTA, as shown in Table 11.

2.4 Training Objectives

To achieve both accurate anomaly detection and precise localization, we jointly optimize image-level classification and pixel-level segmentation objectives. Joint training allows the two tasks to reinforce each other, as global context aids localization while fine-grained cues improve classification. All training is performed on the auxiliary dataset $\mathcal{D}_{aux}$. Given the aggregated text embedding $T_{(\cdot)}^{fus}$ and the global visual feature x_{cls}, we compute the anomaly score and corresponding image-level loss:

$$\hat{y} = \sigma\big(\text{sim}(x_{\text{cls}}, T_{(\cdot)}^{\text{fus}})\big), \quad \mathcal{L}_{\text{image}} = \text{BCE}(\hat{y}, y) \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\text{sim}(\cdot,\cdot)$ is cosine similarity, and BCE is the binary cross-entropy loss comparing the predicted score $\hat{y}$ to the ground-truth image label $y \in \{0,1\}$. For pixel-level localization, following [47], we extract patch features $\{X^{(m)} \in \mathbb{R}^{H\times W\times d}\}_{m\in\mathcal{M}}$ from a set $\mathcal{M}$ of selected layers in f_{visual}. Each $X^{(m)}$ is formed by reshaping the patch token sequence into a spatial grid of size $H \times W$ with d-dimensional embeddings. For each $m \in \mathcal{M}$, we compute an anomaly map using the aggregated text embedding:

$$\hat{G}^{(m)} = \sigma\big(\text{sim}(X^{(m)}, T_{(\cdot)}^{\text{fus}})\big), \quad \mathcal{L}_{\text{pixel}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\big[\text{Focal}(\hat{G}^{(m)}, G) + \text{Dice}(\hat{G}^{(m)}, G)\big] \tag{7}$$

where $\hat{G}^{(m)} \in [0,1]^{H\times W}$ is the predicted anomaly probability map for layer s, G is the ground-truth pixel-level mask, $\text{Focal}(\cdot,\cdot)$ denotes the focal loss [31], and $\text{Dice}(\cdot,\cdot)$ denotes the Dice loss [40]. This combination balances pixel-wise discrimination and overlap accuracy. The final training objective combines both terms $\mathcal{L}_{\text{PILOT}} = \mathcal{L}_{\text{image}} + \mathcal{L}_{\text{pixel}}$.

2.5 Test-Time Adaptation

A central challenge for robust ZSAD, as outlined in the Introduction, is achieving reliable adaptation to target domains that may differ significantly from the auxiliary dataset. While fine-tuning on $\mathcal{D}_{\text{aux}}$ can calibrate model representations, substantial domain shifts often cause conventional approaches to overfit or fail to generalize, especially in the absence of labeled data from the target domain. To overcome these obstacles, we propose a TTA strategy that enables the model to adjust to new domains using only unlabeled target data, without compromising the integrity of previously learned representations.

The core idea of our TTA strategy is to identify a subset of highly reliable predictions (high-confidence pseudo-labels) from the target dataset and leverage these predictions to update model parameters. Specifically, for each image I_t in the unlabeled target dataset $\mathcal{D}$tar, we compute the anomaly score as $\tilde{y}t = \sigma\big(\text{sim}(xcls^t, , T(\cdot)^{\text{fus}})\big)$, where x_{cls}^t is the CLIP visual feature for I_t and $T_{(\cdot)}^{\text{fus}}$ denotes the fused prompt embedding selected and aggregated for I_t using model parameters trained on $\mathcal{D}_{\text{aux}}$. We then identify the most reliable samples by selecting the top and bottom ρ-fractions of scores:

$$A = \text{Top}_{\lceil\rho B\rceil}\{\tilde{y}_t\}_{t=1}^{B}, \qquad N = \text{Bot}_{\lceil\rho B\rceil}\{\tilde{y}_t\}_{t=1}^{B}, \tag{8}$$

where $B = |\mathcal{D}_{\text{tar}}|$ and $0 < \rho < 0.5$. Pseudo-labels are then defined by $\hat{y}_t = 1$ if $t \in A$, and $\hat{y}_t = 0$ if $t \in N$. Adaptation is performed by optimizing an image-level binary cross-entropy loss on the selected samples, i.e., $\mathcal{L}_{\text{image}} = \text{BCE}(\hat{y}_i, \tilde{y}_i)$ for $i \in A \cup N$.

Critically, during TTA, we update only the parameters associated with the learnable prompt pool, including the learnable prompts, their corresponding embeddings s_k and r_k. All other model parameters, such as those in the attribute memory bank and the core CLIP architecture, remain fixed. The updates for each learnable prompt are weighted by the attention scores defined previously (see Section 2.1), meaning prompts with higher attention scores for a given input contribute more significantly to parameter updates. We limit adaptation strictly to the image-level objective because, as shown in our experiments (Section 3.2), pseudo-labels derived at the image level exhibit considerably higher reliability compared to those at the pixel level. Pixel-level pseudo-labeling can be adversely affected by spatial ambiguity and noise exacerbated by domain shifts. By focusing adaptation solely on image-level labels, we reduce the risk of propagating noisy pseudo-labels, ensuring stable, unsupervised adaptation in the target domain. Our experimental results demonstrate that applying such a strategy improves both pixel-level and image-level detection on the target domain.

3 Experiments

Benchmarks. To demonstrate the effectiveness of our method, we conduct experiments on 13 widely used benchmarks spanning industrial inspection and medical imaging. Specifically, we evaluate on six industrial datasets: BTAD [34], VisA [48], MVTec AD [2], DAGM [44], SDD [42], and DTD-Synthetic [1] and seven medical datasets, subdivided into detection-only (HeadCT [25], BrainMRI [22], Br35H [17]) and localization (ISIC [12], CVC ColonDB [43], CVC ClinicDB [4], Kvasir [21]). The medical datasets for anomaly localization contain only anomalous samples and lack normal data; since our TTA strategy requires normal samples, results are reported without TTA (see Appendix H). To ensure strict zero-shot evaluation, auxiliary datasets exclude all target data. More specifically, for a fair comparison, we followed the related literature [8, 47], and used MVTec AD [2] as the auxiliary dataset when reporting the performance on all other datasets, while when reporting the performance on MVTec AD, we trained the model on VisA [48] as the auxiliary dataset.

Baseline. We compare PILOT with the most recent SOTA methods, including SAA[7], CoOp [46], WinCLIP [20], APRIL-GAN [10], AnomalyCLIP [47], and AdaCLIP [8].

Evaluation Metrics. Following standard practices in zero-shot anomaly detection literature such as [10, 47], we evaluated our method using the AUROC. Additionally, we report APs for anomaly detection and use AUPRO [3] to assess pixel-level anomaly localization performance. Results are provided at both dataset-level and domain-level averages.

3.1 Comparison with State-of-the-Art

As reported in Table 1, PILOT establishes state-of-the-art performance across all benchmarks for both image-level anomaly detection and pixel-level localization in the zero-shot setting. This can be benefited from PILOT comprehensive approach in addressing critical challenges prevalent in ZSAD, that explicitly tackles domain shift in test phase. Prior approaches [7, 20] rely solely on predefined prompts, which may not generalize well to unseen domains, whereas methods such as [8, 10, 46, 47] employ lightweight visual adapters or a single learnable prompt to capture anomaly cues, but often overfit to the auxiliary training set, limiting robustness under distribution shift. PILOT distinguishes itself by dynamically identifying the most relevant prompts for each input and refining only those learnable prompt parameters during inference using high-confidence pseudo-labels from the unlabeled target data. Note that while existing baselines can utilize the proposed image-level TTA strategies, these often degrade pixel-level localization performance and do not yield consistent improvements across both detection and localization tasks. On the contrary, PILOT uniquely achieves gains in both image-level and pixel-level metrics under TTA, a phenomenon analyzed in detail in Section 3.2. We also compare against state-of-the-art full-shot methods in Appendix F.1, which show that PILOT achieves image-level accuracy on par.

Table 1: Model performance on public benmarks. The top section reports image-level results (AUROC, AP), and the bottom section reports pixel-level results (AUROC, AUPRO). Best entries are marked with bold and second-best entries with underline.

Dataset	SAA	CoCoOp	WinCLIP	APRIL-GAN	AnomalyCLIP	AdaCLIP	PILOT
Industrial Image-level Anomaly Detection (AUROC, AP)							
MVTec AD	(64.2, 87.0)	(88.2, 94.5)	(91.9, 96.4)	(86.5, 93.2)	(91.6, 96.3)	(91.3, 95.8)	(**92.1**, **96.6**)
VisA	(67.5, 76.2)	(63.2, 68.5)	(78.5, 81.4)	(78.2, 81.3)	(81.8, 85.1)	(81.6, 84.1)	(**84.1**, **85.7**)
BTAD	(59.4, 89.2)	(67.0, 77.6)	(68.4, 71.1)	(73.8, 68.9)	(83.8, 88.4)	(90.4, 90.7)	(**95.3**, **96.8**)
SDD	(68.9, 37.9)	(75.1, 65.3)	(84.5, 77.6)	(80.0, 71.7)	(85.5, 81.7)	(81.4, 72.8)	(**87.9**, **85.7**)
DAGM	(87.3, 88.6)	(87.7, 74.8)	(91.9, 79.7)	(94.6, 84.0)	(96.1, 90.2)	(97.3, **93.1**)	(**98.5**, 92.8)
DTD-Synthetic	(94.6, 93.7)	(92.4, 94.8)	(93.4, 92.8)	(86.7, 95.2)	(93.7, 97.3)	(97.5, 98.2)	(**97.7**, **98.4**)
Mean	(73.7, 78.8)	(78.9, 79.3)	(84.8, 83.2)	(83.3, 82.4)	(88.8, 89.8)	(89.9, 89.1)	(**92.6**, **92.7**)
Medical Image-level Anomaly Detection (AUROC, AP)							
HeadCT	(46.8, 68.0)	(78.4, 78.8)	(81.8, 80.2)	(89.1, 89.4)	(91.0, 92.3)	(86.5, 81.3)	(**94.5**, **96.0**)
BrainMRI	(34.4, 76.7)	(61.3, 44.9)	(86.6, 91.5)	(89.3, 90.9)	(93.2, 94.7)	(93.2, 89.7)	(**97.3**, **97.9**)
Br35H	(33.2, 67.3)	(86.0, 87.5)	(80.5, 82.2)	(93.1, 92.9)	(95.3, 96.0)	(92.1, 90.9)	(**96.7**, **97.7**)
Mean	(38.1, 70.7)	(75.2, 70.4)	(83.0, 84.6)	(90.5, 91.1)	(93.2, 94.3)	(90.6, 87.3)	(**96.2**, **97.2**)
Industrial Pixel-level Anomaly Localization (AUROC, AUPRO)							
MVTec AD	(75.5, 38.1)	(44.4, 11.1)	(85.1, 64.6)	(87.6, 44.0)	(**90.7**, 78.7)	(89.4, 37.8)	(90.3, **80.2**)
VisA	(76.5, 31.6)	(42.1, 12.2)	(79.6, 56.8)	(94.2, 86.8)	(95.3, 85.1)	(95.5, 77.8)	(**96.0**, **87.2**)
BTAD	(65.8, 14.8)	(28.1, 6.5)	(72.7, 27.3)	(60.8, 25.0)	(94.4, 73.6)	(94.8, 32.5)	(**96.3**, **74.9**)
SDD	(78.8, 6.6)	(24.4, 8.3)	(68.8, 24.2)	(79.8, 65.1)	(90.7, 66.6)	(71.7, 17.6)	(**92.4**, **73.4**)
DAGM	(62.7, 32.6)	(17.5, 2.1)	(87.6, 65.7)	(82.4, 66.2)	(95.6, 91.0)	(91.1, 62.3)	(**96.0**, **91.4**)
DTD-Synthetic	(76.7, 60.6)	(14.8, 3.0)	(83.9, 57.8)	(95.3, 86.9)	(97.8, **91.1**)	(**98.2**, 86.2)	(98.0, 90.2)
Mean	(72.7, 30.7)	(28.6, 7.2)	(79.6, 49.4)	(83.4, 62.3)	(94.1, 81.0)	(90.1, 52.4)	(**94.8**, **82.9**)

3.2 Effect of Test-Time Adaptation

Table 2 presents a comparison of all methods evaluated under a unified TTA protocol, where each method adapts learnable parameters across the test set using high-confidence pseudo-labels following PILOT's label-free strategy. This fair setup reveals a key limitation of image-level TTA strategies in existing ZSAD models: while they can yield modest improvements in detection, they often destabilize pixel-level localization due to indiscriminate parameter updates guided by global objectives. This is particularly evident in methods like AnomalyCLIP, which use a single learnable prompt across all test images. In these cases, high-confidence pseudo-labels are often insufficient to represent the diversity of anomaly patterns in the target domain, resulting in imbalanced adaptation (see Appendix F.2) and degraded fine-grained localization. Thus, image-level TTA alone struggles to maintain the alignment between visual and semantic cues necessary for precise localization, especially in complex scenarios. AAnother significant limitation is conventional TTA methods' susceptibility to noisy pseudo-labels, which propagate errors and degrade both detection and localization performance. PILOT's adaptive prompt weighting addresses this by localizing the impact of label noise—updates concentrate on prompts closely aligned with each input, thus minimizing negative effects on the overall prompt pool.

3.3 Module Ablation

Table 3 evaluates individual components of PILOT across industrial datasets, highlighting how our design stabilizes pixel-level predictions during TTA. Using a single prompt, similar

Table 2: Comparison of PILOT vs. state-of-the-art on industrial benchmarks. **Img** and **Pix** denote image-level and pixel-level metrics, respectively. "w/o TTA" and "w/ TTA" indicate performance without and with TTA, respectively. Arrows indicate change in performance: $\uparrow$(increase > 0.5), $\downarrow$(decrease < -0.5), and $\circ$(no significant change, ± 0.5).

Dataset		APRIL-GAN		AnomalyCLIP		AdaCLIP		PILOT	
		w/o TTA	w/ TTA	w/o TTA	w/ TTA	w/o TTA	w/ TTA	w/o TTA	w/ TTA
MVTec AD	Img	(86.5, 93.2)	(82.9↓, 89.8↓)	(91.6, 96.3)	(89.2↓, 95.0↓)	(91.3, 95.8)	(90.1↓, 95.8∘)	(90.2, 95.7)	(92.1↑, 96.6↑)
	Pix	(87.6, 44.0)	(86.0↓, 42.0↓)	(90.7, 78.7)	(88.1↓, 73.5↓)	(89.4, 37.8)	(87.0↓, 29.5↓)	(89.4, 79.2)	(90.3↑, 80.2↑)
VisA	Img	(78.2, 81.3)	(84.5↑, 88.0↑)	(81.8, 85.1)	(84.3↑, 85.8↑)	(81.6, 84.1)	(83.0↑, 85.0↑)	(82.0, 84.1)	(84.1↑, 85.7↑)
	Pix	(94.2, 86.8)	(92.0↓, 84.0↓)	(95.3, 85.1)	(85.5↓, 82.6↓)	(95.5, 77.8)	(93.8↓, 72.6↓)	(94.8, 85.6)	(96.0↑, 87.2↑)
BTAD	Img	(73.8, 68.9)	(76.9↑, 73.2↑)	(83.8, 88.4)	(85.7↑, 93.0↑)	(90.4, 90.7)	(91.6↑, 92.5↑)	(90.0, 91.6)	(95.3↑, 96.8↑)
	Pix	(60.8, 25.0)	(58.0↓, 23.0↓)	(94.4, 73.6)	(94.1∘, 72.0↓)	(94.8, 32.5)	(93.0↓, 25.5↓)	(94.9, 73.0)	(96.3↑, 74.9↑)
SDD	Img	(80.0, 71.7)	(79.9∘, 71.6∘)	(85.5, 81.7)	(85.7∘, 81.8∘)	(81.4, 72.8)	(83.0↑, 74.0↑)	(87.3, 85.2)	(87.9↑, 85.7↑)
	Pix	(79.8, 65.1)	(77.5↓, 63.0↓)	(90.7, 66.6)	(89.9↓, 63.8↓)	(71.7, 17.6)	(66.7↓, 12.5↓)	(88.7, 69.6)	(92.4↑, 73.4↑)
DAGM	Img	(94.6, 84.0)	(95.0∘, 84.5∘)	(96.1, 90.2)	(97.3↑, 92.0↑)	(97.3, 93.1)	(96.8∘, 92.2∘)	(97.1, 91.8)	(98.5↑, 92.8↑)
	Pix	(82.4, 66.2)	(80.5↓, 64.0↓)	(95.6, 91.0)	(93.1↓, 90.2↓)	(91.1, 62.3)	(90.2↓, 58.8↓)	(95.8, 91.8)	(96.0∘, 91.4∘)
DTD-Synthetic	Img	(86.7, 95.2)	(85.0↓, 93.5↓)	(93.7, 97.3)	(93.2∘, 96.9∘)	(97.5, 98.2)	(97.5∘, 98.4∘)	(95.3, 98.0)	(97.7↑, 98.4↑)
	Pix	(95.3, 86.9)	(93.5↓, 84.5↓)	(97.8, 91.1)	(93.3↓, 83.4↓)	(98.2, 86.5)	(94.9↓, 73.3↓)	(97.3, 89.9)	(98.0↑, 90.2∘)

Table 3: Module ablation for PILOT on industrial datasets. **SLP**: Single Learnable Prompt; **+Attr**: SLP + Attribute Memory Bank; **Full**: SLP + Attr + Query & Prompt Pool (Full model). Metrics: Image = (AUROC, AP), Pixel = (AUROC, AUPRO).

	Image-Level		Pixel-Level	
	w/o TTA	w/ TTA	w/o TTA	w/ TTA
SLP	(90.8, 91.6)	(91.4↑, 91.9∘)	(94.2, 82.1)	(90.7↓, 79.8↓)
+Attr	(90.5, 91.3)	(91.2↑, 91.7∘)	(93.9, 81.5)	(94.0∘, 81.5∘)
Full	(90.3, 91.1)	(92.6↑, 92.7↑)	(93.5, 81.5)	(94.8↑, 82.9↑)

to existing ZSAD methods, improves detection but severely degrades localization accuracy due to the lack of spatial supervision, causing visual-textual misalignment. Introducing the fixed attribute memory bank addresses this by providing stable semantic anchors, constraining model adaptation and preserving localization quality through parameter-free orthogonal projection. The complete PILOT framework integrates prompts from the learnable pool and attribute memory bank through a dual-branch fusion mechanism, dynamically weighted by adaptive query functions. During TTA, parameter updates are guided by alignment scores, ensuring prompts aligned with high-confidence samples undergo more substantial refinement. This adaptive update strategy mitigates negative transfer risks from unseen data, enhancing both image-level detection and pixel-level localization.

4 Conclusion

We have introduced PILOT, a ZSAD framework designed to address domain shifts by adaptively integrating anomaly cues through a dual-branch prompt aggregation mechanism, which combines a learnable prompt pool with semantic attribute embeddings. Our image-level TTA strategy further refines the prompt parameters using high-confidence pseudo-labels from unlabeled target data. Extensive experiments on 13 public benchmarks demonstrate that PILOT consistently achieves state-of-the-art performance in both anomaly detection and localization. Future work may explore advanced TTA strategies to jointly improve pixel-level localization and image-level detection, further enhancing ZSAD robustness.

References

[1] 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*, pages 5564–5572, 2023.

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

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

[4] 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.

[5] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5542–5552, 2024.

[6] Jakob Bozic, Domen Tabernik, and Danijel Skocaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021.

[7] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

[8] 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*, pages 55–72. Springer, 2024.

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[10] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: 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.

[11] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*, pages 17–33. Springer, 2024.

[12] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[13] Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2(5), 2023.

[14] 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, pages 6568–6576, 2022.

[15] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Fine-tuning clip's last visual projector: A few-shot cornucopia. *arXiv preprint arXiv:2410.05270*, 2024.

[16] Ruoyu Feng, Tao Yu, Xin Jin, Xiaoyuan Yu, Lei Xiao, and Zhibo Chen. Rethinking domain adaptation and generalization in the era of clip. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 2585–2591. IEEE, 2024.

[17] A. Hamada. Br35h: Brain tumor detection 2020. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection, 2020. Accessed: [insert access date].

[18] Hadi Hojjati and Narges Armanfard. Dasvdd: Deep autoencoding support vector data descriptor for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):3739–3750, 2023.

[19] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

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

[21] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pal Halvorsen, Thomas De Lange, Dag Johansen, and Havard 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*, pages 451–462. Springer, 2020.

[22] Pranita Balaji Kanade and P. P. Gumaste. Brain tumor detection using mri images, 2015.

[23] Ali Karami, Thi Kieu Khanh Ho, and Narges Armanfard. Graph-jigsaw conditioned diffusion model for skeleton-based video anomaly detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4237–4247. IEEE, 2025.

[24] 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*, pages 19113–19122, 2023.

[25] F.C. Kitamura. Headct-hemorrhage. https://www.kaggle.com/dsv/152137, 2018.

[26] Thomas Lai, Thi Kieu Khanh Ho, and Narges Armanfard. Open-set multivariate time-series anomaly detection. *arXiv preprint arXiv:2310.12294*, 2023.

[27] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[28] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[29] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020.

[30] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

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

[32] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

[33] Aldo Marzullo and Marta Bianca Maria Ranzini. Exploring zero-shot anomaly detection with clip in medical imaging: Are we there yet? *arXiv preprint arXiv:2411.09310*, 2024.

[34] 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)*, pages 01–06. IEEE, 2021.

[35] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

[36] Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146. URL https://aclanthology.org/P19-1146/.

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

[38] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022.

[39] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

[40] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[41] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[42] Domen Tabernik, Samo Sela, Jure Skvarc, and Danijel Skocaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

[43] 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.

[44] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, volume 6, page 11, 2007.

[45] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 11772–11781, 2021.

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[47] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=buC4E91xZE.

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