

Defect Cue-Preserved Structural Feature Refinement for Few-Shot Anomaly Detection

Le Jiang¹, Yan Huang^{1,*}, Zhen Xu¹, Yong Xu^{1,2,3}, Hau-San Wong⁴, Si Wu^{1,*}

¹South China University of Technology, China

²Guangdong Provincial Key Laboratory of Multimodal Big Data Intelligent Analysis, China

³State Key Laboratory of Advanced Papermaking and Paper-based Materials, China

⁴City University of Hong Kong, China

{csjiangle@mail., aihuangy@, csxuzhen@mail., yxu@}scut.edu.cn,
 cshswong@cityu.edu.hk, cswusi@scut.edu.cn

Abstract

Modern industrial quality control heavily relies on automated anomaly detection. While few-shot anomaly detection addresses the challenge of limited labeled data, real-world inspection faces a vast diversity of anomaly types, sizes, and shapes. We identify the primary cause for the anomaly detection difficulty as the progressive loss of detect cues as they pass through deep feature extraction pipelines. To counteract the defect cue fading, we propose a Defect Cue-Preserved Structural Feature Refinement model, referred to as DCP-SFR. Recognizing that early-stage cues are paramount, we design a conditional anomaly cue amplification module to produce an initial anomaly score map, which is then enhanced to increase the contrast between anomalous and normal regions. The amplified cues is subsequently used for reconstruction-based anomaly localization, by anchoring attention on true anomaly regions to preserve spatial integrity and prevent drift. Further, we incorporate a structure-aware segmentation refinement stage to improve anomaly segmentation in terms of edge alignment, thereby significantly improve boundary accuracy. On the MVTec AD and VisA benchmarks, DCP-SFR achieves state-of-the-art performance, with an image-level AUROC of 97.3% and a pixel-level AUROC of 98.2%, demonstrating strong cross-domain generalization performance.

1. Introduction

Automated anomaly detection (AD) is a foundational technology for controlling industrial product quality. Conventional AD is often constrained by a reliance on substantial annotated data [15, 21, 22, 30, 32, 37, 39, 40, 43]. While

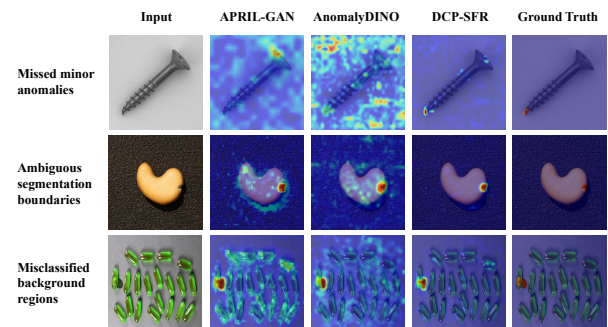


Figure 1. Image examples of the three key FSAD challenges. Compared with the leading FSAD models, the proposed DCP-SFR achieves superior performance in focusing more on anomalous regions and localizing defect boundaries.

zero-shot anomaly detection (ZSAD) removes such a dependency [1, 4, 6, 12, 23, 28, 47], an absence of reference images makes it difficult to reliably localize defects that lack strong, explicit visual cues. By contrast, few-shot anomaly detection (FSAD) offers a more practical solution, leveraging a small set of reference images to enable rapid deployment across diverse products [13, 16, 17, 35, 48]. The main challenge for real-world FSAD lies in contending with a wide variety of defect types, scales and unseen patterns. The variety and complexity of defects underscores the critical importance of subtle, anomaly-related cues for accurate detection. On the other hand, the signal loss inherent in multi-stage feature extraction pipelines leads to defect cue fading, a phenomenon manifesting as mislocalized anomalous regions and compromised boundary precision. Therefore, preserving the vital defect cues is of great significance to precise anomaly detection.

Existing FSAD models can be broadly categorized into three categories. The methods of the first category aim to model the normality, often through a reconstruction strat-

*Corresponding authors.

egy, like FastRecon [13]. Reconstruction-based models detect anomalies as the regions with high reconstruction error but can misclassify normal samples in complex scenes due to limited reference coverage. The methods of the second category are developed to synthesize anomalies to augment training data, such as SPD [48], AnomalyDiffusion [16] and AnomalyAny [35]. However, synthetic defects are always accompanied by the domain discrepancy that confuses defect representation learning. The methods of the third category perform the comparison between the features extracted from the local regions of a target image and normal references. A critical prerequisite for a meaningful comparison is ensuring that the features are spatially aligned, as addressed by early methods like RegAD [17] which focused on feature-level alignment. Subsequently, the focus of research shifted toward leveraging powerful pretrained models, such as CLIP [33] and DINOv2 [31], to extract rich semantic and structural features. State-of-the-art methods like PatchCore [34], AnomalyGPT [14], WinCLIP [18], and AnomalyDINO [8] operate by matching the features in a well-established embedding space to compute discrepancy scores. However, there is a lack of effective mechanism to counteract the progressive fading of defect cues during processing, which significantly limits the reliability of existing approaches, as shown in Figure 1.

In this work, we propose DCP-SFR, a framework designed to counteract progressive defect cue fading. The proposed approach is built on a multi-stage information preservation strategy, aiming to consistently enhance defect-relevant structural information from initial extraction to final prediction. Specifically, we incorporate a conditional anomaly cue amplification module, which leverages multi-level features and learnable prompts to perform an initial differentiation between normal and anomalous regions. The resulting prediction scores are then amplified to produce high-contrast anomaly cues that robustly preserves initial defect indication. Guided by the amplified cues, anomaly cue-guided localization module achieves more precise defect localization. By reconstructing the query features from normal references, the resulting residual map, derived from the difference between the original and reconstructed features, effectively pinpoints defect locations. Structure-aware segmentation refinement module by fusing the localization map with a regularization map computed from visual feature similarity. A dedicated edge alignment regularization is imposed to ensure that the segmentation result has sharp and precise boundaries. Extensive experiments on the MVTec AD and VisA datasets demonstrate that DCP-SFR is able to achieve superior performance. The main contributions of this work are summarized as follows:

- To address defect cue fading in FSAD, a problem largely overlooked by prior works, we propose DCP-SFR to preserve and refine defect cues throughout the pipeline, lead-

ing to superior performance in both segmentation and image-level classification.

- To preserve defect cues, we design a conditional amplification module to enhance subtle initial signals, and the amplified cues are leveraged to guide defect localization, achieving high-precision anchoring on anomalous regions and effectively preventing drift.
- We further impose structural alignment regularization, which significantly enhances the contour fidelity of anomaly segmentation, leading to the superior performance of DCP-SFR over the existing FSAD methods.

2. Related Work

To provide an overview of FSAD, we categorize existing methods based on the adopted strategies: (1) reconstruction-based methods, (2) anomaly generation methods, and (3) feature comparison-based methods.

2.1. Reconstruction-based Methods

There are a number of FSAD methods [41, 42], which operate on a core assumption that the representation learnt exclusively from normal samples can be used to exactly reconstruct the normal regions but will fail to reconstruct anomalous ones. Anomaly detection is performed by subtracting the reconstructed image from the query image, and the resulting significant discrepancies indicate defects. FastRecon [13] leverages a limited amount of normal references to reconstruct query images via a specialized regression model with distribution regularization, and identity defects according to reconstruction errors. However, due to the limited references, the feature space learned by the reconstruction-based methods provides inadequate coverage of normal patterns, leading to a high rate of false positives.

2.2. Anomaly Generation Methods

To address the data imbalance inherent in anomaly detection, another category of FSAD methods focuses on augmenting training datasets by generating synthetic anomalies. Early approaches relied on manually crafted defects. For instance, SPD [48] performs self-supervised contrastive learning by using non-generative anomalies, while Wang et al. [36] proposed an energy-based model that reconstructs the images with manually injected random noise.

A significant drawback of handcrafted anomalies is the lack of realism. To overcome the issue, recent works leverage powerful generative models. AnomalyDiffusion [16] and AnomalyAny [35] utilize diffusion models to generate defects that more closely match real-world distributions, thereby partially alleviating the scarcity of real anomaly samples. However, a challenge for the above generative methods is the potential domain gap between synthetic and real defects, which can cause the decision boundaries learnt by the models deviate from the ground-truth regions.

2.3. Feature Comparison-based Methods

Feature comparison-based methods detect anomalies by measuring the similarity between features extracted from query samples and those of normal references, leveraging multi-scale or semantic correlations to highlight deviations from learned normal patterns. SPADE [7], for instance, performs a training-free kNN-based [11] feature comparison between query and reference images, where kNN retrieves the most similar features to assess normality. PaDiM [9] models the normal feature distribution with multivariate Gaussians from pretrained CNN embeddings and incorporates inter-level correlations to improve defect localization. However, spatial misalignment between query and reference features can hinder accurate anomaly detection. To address spatial misalignment, RegAD [17] introduces a generalizable feature alignment model, alleviating the misalignment problem. Several more recent and influential methods focus on memory-bank-based feature matching. PatchCore [34] constructs a coreset of representative normal patch features for efficient similarity scoring. Building on the memory bank, GraphCore [38] aims to learn a visual isometric invariant features as an anomaly measurement feature for more precise detection.

Another research line leverages the pre-trained cross-modal model like CLIP [33]. WinCLIP [18] implemented a training-free, multi-scale feature comparison between query and reference images, demonstrating the potential of vision-language models for industrial inspection. Li et al. [26] proposed a transformation to warp query features closer to the normal manifold, improving similarity matching.

In addition, it is an effective way to fine-tune or adapt CLIP for FSAD, typically by learning effective textual prompts or image patch representations. APRIL-GAN [5] remaps image features into a joint embedding space to improve text-image alignment. Similarly, SimCLIP [10] incorporates adapters and prompt tuning to enhance cross-modal synergy for effective ZSAD. SimCLIP is further extended to FSAD in SimCLIP+ via relational modeling over visual embeddings. On the other hand, PromptAD [25] constructs a large number of negative samples used to guide prompt learning in one-class setting. InCTRL [46] trains a generalist anomaly detection model using in-context residual learning with normal prompts, enabling cross-domain anomaly detection without target-specific fine-tuning.

Beyond the CLIP-centric paradigm, researchers have also leveraged other powerful pretrained Large Language Models (LLMs) for FSAD. AnomalyGPT [14] couples visual encoders with LLMs to allow for multi-turn interactive querying for defect analysis. ADFormer [45], which fuses CNN features with transformer layers to capture both local textures and long-range dependencies. The DINOv2 [31] model has emerged as a popular visual backbone, and AnomalyDINO [8] directly leverages it for feature compar-

ison. LogSAD [44] adopts a hybrid strategy to integrate DINOv2 with the Segment Anything Model (SAM) [20] to enhance final mask prediction. INP-Former [29] extracts an intrinsic normal prototype from the query image and reconstructs only normal tokens, with reconstruction errors serving as the anomaly indicator.

Difference from the existing methods: The above methods treat few-shot anomaly detection as a static, one-shot matching problem, and overlook the progressive nature of defect cue fading within deep feature extraction pipelines. Compared with prevalent feature-matching methods that compute discrepancy scores with a fixed feature representation, we propose DCP-SFR, which leverages dynamic anomaly cues to refine the entire detection process. Toward this end, we design a conditional amplification module to enhance weak defect signals, and then use the resulting cues to guide a reconstruction-based localization process.

3. Methodology

To address the progressive fading of anomaly signals across the detection pipeline, we propose DCP-SFR, a unified framework for few-shot anomaly detection, as illustrated in Figure 2. DCP-SFR is designed to address the two subtasks: image segmentation and anomaly segmentation. There are a Conditional Anomaly Cue Amplification module (CACA) and an Anomaly Cue-guided Localization module (ACL), built on top of pretrained models, which form a shared processing core that is applied to both subtasks. For the segmentation task, a Structure-aware Segmentation Refinement module (SSR) is appended to enhance boundary precision. The primary distinction lies in the features they process: the segmentation task operates on spatial patch features that preserve locational information, while the classification task operates on the global class token that aggregates overall image context. We will first detail the core modules from the perspective of the segmentation task (Sec. 3.1-3.3). Subsequently, we will describe how the classification task leverages a parallel processing stream with a similar logic and conclude with the overall model optimization strategy (Sec. 3.4).

3.1. Conditional Anomaly Cue Amplification

A primary challenge in AD is the fading of subtle defect cues, which are often overshadowed by background noise. To tackle the issue, we design an effective module to conditionally amplify potential anomaly signals based on the semantic guidance from learnable prompts, preserving subtle defect cues from early feature extraction layers.

Specifically, we leverage two pretrained models as backbones: a CLIP [33] model, comprising an image encoder E_I and a text encoder E_T ; and a DINOv2 [31] image encoder E_D . Given a query image x and K reference normal images $\{x'_k | k = 1, \dots, K\}$, we extract features using these

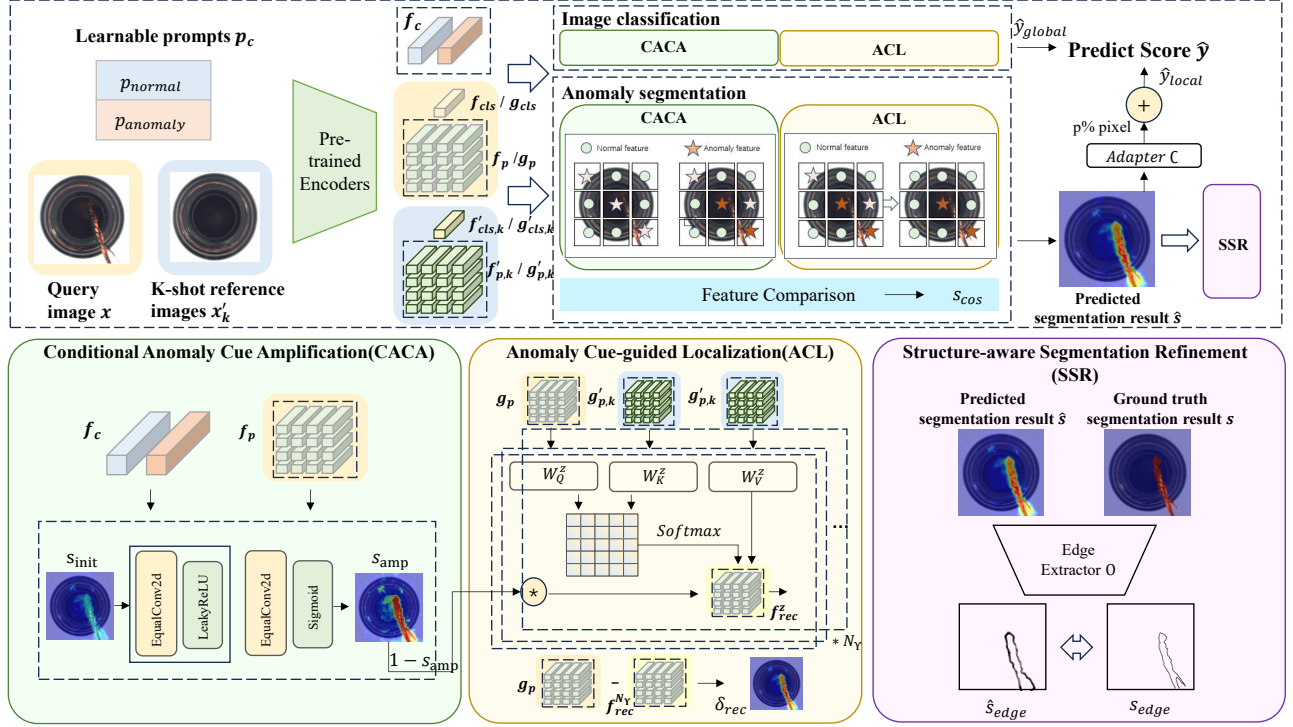


Figure 2. Overview of the DCP-SFR framework. The architecture comprises a CACA Module and an ACL Module, forming a shared core built on pretrained features for both segmentation and classification subtasks. Besides, the SSR maintains geometric accuracy and edge sharpness of segmentation result.

backbones. To preserve early-stage defect cues that might fade in deeper layers, we extract patch features from multiple intermediate layers of E_I :

$$\begin{aligned} \{f_{cls}, f_p\} &= E_I(x), \\ \{f'_{cls,k}, f'_{p,k}\} &= E_I(x'_k), \end{aligned} \quad (1)$$

where f_{cls} is the class token and f_p is a set of multi-level patch features from the query image x . Specifically, f_p is denoted as $\{f_{p,j} | j = 1, \dots, m\}$, where $m = 4$ and each element $f_{p,j}$ corresponds to the features from the 6th, 12th, 18th, and 24th layers, respectively. Analogously, $f'_{cls,k}$ denotes the class token from the k -th reference image x'_k , and $f'_{p,k}$ represents the set of patch features, $\{f'_{p,k,j} | j = 1, \dots, m\}$. In addition, we adopt E_D to extract features that are more attentive to fine-grained details:

$$\begin{aligned} \{g_{cls}, g_p\} &= E_D(x), \\ \{g'_{cls,k}, g'_{p,k}\} &= E_D(x'_k), \end{aligned} \quad (2)$$

where g_{cls} and g_p are the class token and patch features from the query image, $g'_{cls,k}$ and $g'_{p,k}$ are the corresponding features from the reference images.

We utilize learnable prompts to guide our model toward precise anomaly detection. We define the set of class categories as $\mathcal{C} = \{\text{normal}, \text{anomaly}\}$. For each category $c \in \mathcal{C}$, we design a corresponding learnable prompt p_c with the following structure:

$$\begin{aligned} p_{\text{normal}} &= [U_1][U_2] \dots [U_n], \\ p_{\text{anomaly}} &= [T_1][T_2] \dots [T_n][\text{damaged}], \end{aligned} \quad (3)$$

where $[U_1] \dots [U_n]$ and $[T_1] \dots [T_n]$ are randomly initialized sequences of learnable vectors, and n is the specified length. In this design, we use the word ‘[damaged]’ as a semantic anchor for the anomaly prompt. This anchor provides the model with a generic prior about the concept of anomaly. The associated learnable vectors, $[T_1] \dots [T_n]$, are then responsible for capturing and adapting to the diverse and subtle defect patterns specific to each task during training. The pretrained text encoder E_T maps these class prompts p_c into their corresponding embedding vectors: $f_c = E_T(p_c)$.

We interact the learned prompt embeddings f_c with the preserved multi-level image patch features $f_{p,j}$ to generate an initial anomaly score map, denoted as s_{init} :

$$s_{\text{init}} = \frac{1}{m} \sum_{j=1}^m \frac{\exp(\langle f_{\text{anomaly}}, f_{p,j} \rangle / \tau)}{\sum_{c \in \mathcal{C}} \exp(\langle f_c, f_{p,j} \rangle / \tau)}, \quad (4)$$

where $\langle \cdot, \cdot \rangle$ represents the cosine similarity operation and τ is a temperature hyperparameter. For each patch feature $f_{p,j}$ from the m different layers, the above formula aims to compute the probability of belonging to the anomaly class. s_{init} is obtained by averaging the resulting probability maps from the multiple layers, resulting in a more robust estimation. As a result, we effectively translate the preserved early-stage cues into an initial anomaly map that pinpoints regions semantically aligned with the concept of anomaly, providing critical guidance for the subsequent stages.

To enhance the contrast of the initial anomaly score map

s_{init} , we perform refinement by learning a non-linear transformation A as follows:

$$s_{\text{amp}} = \frac{1}{1 + e^{-A(s_{\text{init}})}}, \quad (5)$$

where A functions to amplify even faint anomaly responses while suppressing the signal values in normal regions. The amplified anomaly score map s_{amp} leads to a clearer separation between anomalous and normal areas. The high-contrast map serves as a critical spatial signal to guide the subsequent module to focus more on high-probability anomaly regions.

3.2. Anomaly Cue-guided Localization

While the previous stage amplifies and highlights potential defect cues, a more specific module is required for precise localization. Our Anomaly Cue-guided Localization module achieves the goal by leveraging the amplified cue map to guide a reconstructive process. The main idea behind the module is to reconstruct an anomaly-free version of the query features, conditioned on the spatial guidance provided by the anomaly cue s_{amp} . By computing the residual between the original features and the reconstructed features, we can precisely identify and localize defects that cannot be reconstructed by normal patterns.

We reconstruct the anomaly-free query features using a module of stacked cross-attention layers. The initial reconstructed feature f_{rec}^0 is generated by using the fine-grained features g_p of the query image as the query:

$$\begin{aligned} f_{\text{rec}}^0 &= \Upsilon(W_Q^0 g_p, W_K^0 \bar{g}_p', W_V^0 \bar{g}_p') \\ &= \text{Softmax} \left(\frac{(W_Q^0 g_p)(W_K^0 \bar{g}_p')^T}{\sqrt{d_k}} \right) (W_V^0 \bar{g}_p'), \end{aligned} \quad (6)$$

where Υ represents the cross-attention operation, W_Q^0 , W_K^0 , and W_V^0 are learnable linear projection matrices for the initial layer, and d_k is the dimension of the key vectors. In the above equation, $\bar{g}_p'$ denotes the mean of all K reference image features $\{g_{p,k}'\}$, providing a stable reference of normal patterns as the key and value.

Next, the f_{rec}^0 is fed into N_r stacked cross-attention blocks for layer-wise refinement. In the z -th layer, the output from the previous layer $f_{\text{rec}}^{(z-1)}$ is spatially modulated by the anomaly cue s_{amp} and then used as the new query:

$$\begin{aligned} \hat{f}_{\text{rec}}^{(z-1)} &= f_{\text{rec}}^{(z-1)} \odot (1 - s_{\text{amp}}), \\ f_{\text{rec}}^z &= \Upsilon(W_Q^z \hat{f}_{\text{rec}}^{(z-1)}, W_K^z \bar{g}_p', W_V^z \bar{g}_p'), \end{aligned} \quad (7)$$

where $\odot$ denotes element-wise multiplication, and the projection matrices W_Q^z , W_K^z , and W_V^z are unique to each refinement layer $z \in \{1, \dots, N_r\}$. The $(1 - s_{\text{amp}})$ suppresses the features in the regions marked by the anomaly cue, compelling the current layer to rely more on the normal reference features $\bar{g}_p'$. After the refinement with N_r stacked layers, the final reconstructed feature $f_{\text{rec}}^{N_r}$ is obtained, and the

reconstruction-based residual map δ_{rec} is then derived by computing the L1 distance between the original and reconstructed features:

$$\delta_{\text{rec}} = \|g_p - f_{\text{rec}}^{N_r}\|_1, \quad (8)$$

The residual map δ_{rec} pinpoints the precise locations of defects. The anomaly cue-guided reconstruction strategy ensures that the localization is not only accurate by learning the deep structure of normal patterns but also robust against background noise.

3.3. Structure-aware Segmentation Refinement

Considering that residual map δ_{rec} may still contain false positives caused by normal texture variations, and exhibit coarse boundaries. We introduce a further refinement strategy designed to simultaneously enhance the structural accuracy of the segmentation result. Toward this end, we perform a non-parametric cosine similarity matching to filter false positives and impose structural alignment regularization to obtain precise anomaly contours.

We treat the feature map g_p as a set of spatial feature vectors (tokens), denoted as $g_p = \{g_p(l)\}_{l=1}^L$, where L denotes the number of vectors. Our strategy involves the comparison of each query token $g_p(l)$ against a comprehensive feature memory bank $\mathcal{M}$, which aggregates all normal patterns from the K reference images as follows:

$$\mathcal{M} = \bigcup_{k=1}^K \{g_{p,k}'(l)\}_{l=1}^L. \quad (9)$$

The comparison yields a similarity map Ψ , where each value represents the maximum cosine similarity between a query token and all the tokens within the memory bank as follows:

$$\Psi(l) = \max_{t \in \mathcal{M}} \langle g_p(l), t \rangle. \quad (10)$$

A regularization map is derived from this similarity map by inversion: $s_{\text{cos}} = 1 - \Psi$, where the regions highly similar to normal patterns will have scores close to zero and are thus effectively suppressed.

We obtain the segmentation result $\hat{s}$ by a weighted fusion of the residual map δ_{rec} and the regularization map s_{cos} :

$$\hat{s} = \delta_{\text{rec}} + \lambda s_{\text{cos}}, \quad (11)$$

where λ is weighting factor. This fusion strategy brings two advantages: δ_{rec} derived from reconstruction residuals, offers high sensitivity in detecting all potential anomalous regions, while at the same time s_{cos} provides high precision by rejecting false positives.

To address the issues of diffuse segmentation boundaries and spurious predictions on isolated background regions, we impose structural alignment regularization to enhance contour fidelity and improve boundary precision, and the corresponding loss function $\mathcal{L}_C$ measures the edge consistency between our final predicted segmentation map $\hat{s}$

and the ground-truth s . We adopt an edge extractor, denoted as $O(\cdot)$, to obtain the edge maps: $s_{\text{edge}} = O(s)$ and $\hat{s}_{\text{edge}} = O(\hat{s})$, and $\mathcal{L}_C$ is defined as follows:

$$\mathcal{L}_C = \mathbb{E}_x \left[1 - \frac{2 \sum_i \hat{s}_i^{\text{edge}} s_i^{\text{edge}} + \zeta}{\sum_i (\hat{s}_i^{\text{edge}})^2 + \sum_i (s_i^{\text{edge}})^2 + \zeta} \right], \quad (12)$$

where ζ is a smoothing factor to prevent division by zero. i indicates the i -th query image in the training data.

The cosine similarity matching strategy reduces false positives in the segmentation result, while the structural alignment enhances boundary fidelity. Minimizing $\mathcal{L}_C$ leads to the final segmentation map $\hat{s}$ becomes reliable in terms of content and structure.

3.4. Model Optimization

Image classification via global-local prediction fusion. To perform image-level binary classification (normal/anomaly), our model combines global information derived from class tokens with local cues from the segmentation map.

Analogous to Eq. (4), the class token f_{cls} is compared with the learned prompt embeddings f_c to obtain an initial global anomaly score. Additionally, a reconstruction module, logically identical to the one in Sec. 3.2, takes the class token g_{cls} and its reference $\bar{g}'_{cls}$ as input to produce a reconstructed token g''_{cls} . The residual, $\eta_{rec} = \|g_{cls} - g''_{cls}\|_1$, captures the deviation of the global features from the normal prototype and is fed into a lightweight adapter network C to generate the final global classification score, $\hat{y}_{\text{global}} = C(\eta_{rec})$.

On the other hand, the model extracts local criteria directly from the final segmentation map, $\hat{s}$. We design another lightweight network, $\kappa(\cdot)$, which takes the segmentation map $\hat{s}$ as input and predicts a sampling percentage, $p = \kappa(\hat{s})$, and we select the top- $p\%$ of pixels from $\hat{s}$. The mean value of these pixel scores constitutes the classification score $\hat{y}_{\text{local}}$ based on local cues. The final image-level classification result is a weighted fusion of the global and local scores as follows:

$$\hat{y} = \hat{y}_{\text{global}} + \mu \hat{y}_{\text{local}}, \quad (13)$$

where μ is weighting factor and $\hat{y}$ represents the predicted label. In the process of model training, we establish clear objectives tailored to anomaly classification, and the cross-entropy loss function $\mathcal{L}_{CE}$ is defined as follows:

$$\mathcal{L}_{CE} = \mathbb{E}_x [-(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))], \quad (14)$$

where y represents the ground truth label.

Anomaly Segmentation. Anomaly areas are typically much smaller than normal areas and we thus adopt a combination of focal loss function $\mathcal{L}_{Focal}$ [27] and dice Loss function $\mathcal{L}_{Dice}$ [24] to optimize the model as follows:

$$\mathcal{L}_{Focal} = \mathbb{E}_x [-\alpha(1 - \hat{s})^\gamma (s \log(\hat{s}) + (1 - \alpha)\hat{s}^\gamma (1 - s) \log(1 - \hat{s}))], \quad (15)$$

where $\hat{s}$ is the final segmentation map, s is the ground truth, γ is the focus factor used to adjust the weight of difficult and easy samples. In the above equation, α is the balance factor used to adjust the influence between positive and negative samples. We also employ $\mathcal{L}_{Dice}$ to directly optimize the spatial overlap of the segmentation masks as follows:

$$\mathcal{L}_{Dice} = \mathbb{E}_x [1 - \frac{2 \sum_i \hat{s}_i s_i + \zeta}{\sum_i \hat{s}_i^2 + \sum_i s_i^2 + \zeta}], \quad (16)$$

where ζ is a very small number used to prevent the denominator from being zero.

By integrating the above loss functions, the optimization formulation of our proposed model is expressed as follows.

$$\min_{\Theta} \mathcal{L}_{CE} + \mathcal{L}_{Focal} + \mathcal{L}_{Dice} + \beta \mathcal{L}_C, \quad (17)$$

where Θ represents all the trainable parameters in our model, β is the weighting factor for $\mathcal{L}_C$. It is worth noting that we train all the trainable components from scratch.

4. Experiments

We conduct comprehensive experiments on MVTec AD [2] and VisA [48] to evaluate the performance of our proposed framework. We begin by detailing the datasets, competitive methods, and implementation details. Subsequently, We present and analyze the main quantitative and qualitative results. Finally, a series of ablation studies are presented to investigate the contributions of individual components.

4.1. Experimental Settings

Datasets. We evaluated the effectiveness of DCP-SFR on two public industrial datasets: MVTec AD and VisA, which contain non-overlapping object categories and distinct defect types. The datasets comprise images representing 27 categories of industrial products, such as candle, bottle, zipper, pill and printed circuit board (PCB). Both datasets provide high-resolution images with pixel-level annotations of common objects, which are 700^2 to 1024^2 for MVTec AD and about $1500 * 1000$ for VisA.

Evaluation metrics and comparison methods. We evaluate the performance of methods on both image-level classification and pixel-level segmentation tasks. For classification, the metrics are the image-level Area Under the Receiver Operating Characteristic curve (i-AUROC), the Area Under the Precision-Recall curve (i-AUPR), and the maximum F1-score (i-F1-max). For segmentation, we report the pixel-wise AUROC (p-AUROC), the Per-Region Overlap score (p-PRO) [3], the pixel-wise maximum F1-score (p-F1-max), and the pixel-level Average Precision (p-AP). We conduct a comparative analysis against representative FSAD methods: WinCLIP [18], APRIL-GAN [5], AnomalyGPT [14], PromptAD [25], InCTRL [46], and Anomaly-DINO [8]. For a fair comparison, the competing methods are implemented using the publicly available source codes.

Implementation details. DCP-SFR is implemented in

Table 1. Quantitative comparison between DCP-SFR and the competing methods on industrial datasets. **Boldface** indicates the best results and underline indicates suboptimal results.

Setup	Method	Public	MVTec AD Classification			MVTec AD Segmentation				VisA Classification			VisA Segmentation			
			i-AUROC	i-AUPR	i-F1-max	p-AUROC	p-PRO	p-F1-max	p-AP	i-AUROC	i-AUPR	i-F1-max	p-AUROC	p-PRO	p-F1-max	p-AP
1-shot	WinCLIP+	CVPR' 2023	93.1	96.5	93.7	95.2	87.1	55.9	47.3	83.8	85.1	83.1	96.4	85.1	41.3	33.0
	APRIL-GAN	CVPR' 2023	92.0	95.8	92.4	95.1	90.6	54.2	51.8	<u>91.2</u>	<u>93.3</u>	<u>86.9</u>	96.0	90.0	38.5	30.9
	AnomalyGPT	AAAI' 2024	94.1	93.7	92.5	95.3	86.9	50.8	47.5	87.4	80.0	80.2	96.2	88.6	37.7	29.1
	PromptAD	CVPR' 2024	94.6	96.1	94.2	95.9	90.1	56.0	53.9	86.9	86.9	81.4	96.7	91.5	40.7	33.1
	InCTRL	CVPR' 2024	94.1	96.9	94.0	-	-	-	-	86.6	87.4	84.4	-	-	-	-
	AnomalyDINO	WACV' 2025	<u>96.6</u>	<u>98.2</u>	<u>95.8</u>	<u>96.8</u>	92.7	<u>60.2</u>	<u>56.5</u>	87.4	89.0	84.3	<u>97.8</u>	<u>92.5</u>	45.1	<u>40.1</u>
	DCP-SFR	-	97.3	98.8	96.7	96.9	92.7	60.7	61.2	92.0	93.7	87.8	98.2	93.3	<u>44.9</u>	40.4
2-shot	WinCLIP+	CVPR' 2023	94.4	97.0	94.4	96.0	88.4	58.4	49.3	84.6	85.8	83.0	96.8	86.2	43.5	34.6
	APRIL-GAN	CVPR' 2023	92.4	96.0	92.6	95.5	91.3	55.9	53.4	<u>92.2</u>	94.2	<u>87.7</u>	96.2	90.1	39.3	31.6
	AnomalyGPT	AAAI' 2024	95.5	94.9	92.8	95.6	88.1	53.1	49.2	88.6	82.1	83.4	96.4	89.9	37.8	29.0
	PromptAD	CVPR' 2024	95.7	96.3	94.1	96.2	90.3	56.4	55.0	88.3	87.5	82.9	97.1	91.8	40.7	34.2
	InCTRL	CVPR' 2024	95.1	97.8	94.8	-	-	-	-	88.0	88.6	86.0	-	-	-	-
	AnomalyDINO	WACV' 2025	<u>96.9</u>	<u>98.2</u>	<u>96.1</u>	<u>97.0</u>	<u>93.1</u>	<u>61.0</u>	<u>61.1</u>	89.7	90.7	86.3	<u>98.0</u>	<u>93.4</u>	<u>46.6</u>	<u>40.3</u>
	DCP-SFR	-	97.7	98.9	97.0	97.1	93.3	61.6	61.7	92.5	94.2	88.0	98.4	93.8	46.9	40.6
4-shot	WinCLIP+	CVPR' 2023	95.2	97.3	94.7	96.2	89.0	59.5	50.7	87.3	88.8	84.2	97.2	87.6	<u>47.0</u>	34.9
	APRIL-GAN	CVPR' 2023	92.8	96.3	92.8	95.9	91.8	56.9	54.5	92.6	<u>94.5</u>	88.4	<u>96.2</u>	90.2	40.0	32.2
	AnomalyGPT	AAAI' 2024	96.3	95.1	92.9	96.2	88.5	54.9	51.1	90.6	82.6	83.7	96.7	90.4	38.4	30.8
	PromptAD	CVPR' 2024	96.6	96.5	94.2	96.5	90.3	56.6	55.3	89.1	87.8	83.6	97.4	91.8	40.9	34.4
	InCTRL	CVPR' 2024	95.1	97.6	95.1	-	-	-	-	89.3	90.4	86.4	-	-	-	-
	AnomalyDINO	WACV' 2025	<u>97.7</u>	<u>98.7</u>	<u>96.6</u>	<u>97.2</u>	<u>93.4</u>	<u>61.8</u>	<u>61.5</u>	<u>92.6</u>	92.9	89.7	<u>98.2</u>	<u>94.1</u>	46.4	<u>40.5</u>
	DCP-SFR	-	98.0	98.9	96.8	97.5	93.5	61.9	61.8	92.8	94.7	88.9	98.6	94.2	47.1	40.6

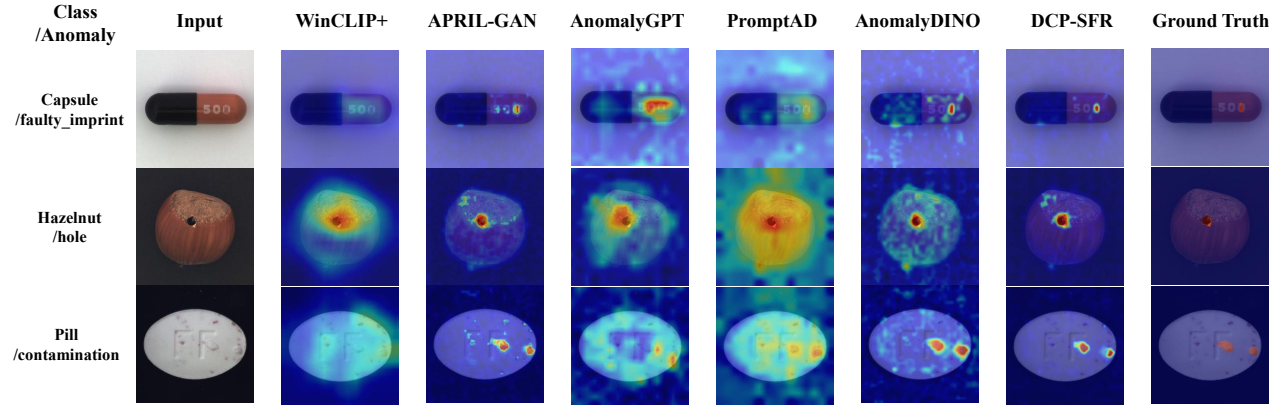


Figure 3. Qualitative comparison of segmentation results between DCP-SFR and the competing methods. The results demonstrate that DCP-SFR accurately detects minor anomalies, minimizes background false positives, and preserves sharp segmentation edges.

PyTorch and trained on a single NVIDIA GeForce RTX 3090. We employ the ViT-L/14@336px variant of the CLIP model and the DINOv2 as backbones, both of which remain frozen during training. The length of the learnable prompt vectors is set to 12, following AnomalyCLIP. The model is trained for 25 epochs using a batch size of 8. We use the Adam optimizer [19] with a learning rate of 0.01. The weighting factors are set as follows: λ in Eq. (11) is 1.0, μ in Eq. (13) is 2.0, α in Eq. (15) is 0.9, and β in Eq. (17) is 2.0. We find that our model converges effectively without heavy tuning. For the cross-domain evaluation, we train the models on MVTec AD and test on VisA, and vice versa.

4.2. Quantitative Comparison

We conducted experiments across 1-, 2-, and 4-shot settings. As shown in Table 1, DCP-SFR demonstrates almost consistent superiority across all evaluated scenarios on both datasets. Even in the challenging 1-shot setting, DCP-SFR maintains exceptional performance. On the MVTec AD segmentation task, our model surpasses AnomalyDINO by about 5 points in terms of AP, showcasing our model’s

robustness with extremely limited references. As the number of reference images increases, the performance of DCP-SFR steadily improves, while at the same time achieving state-of-the-art results across the board. Notably, in the VisA segmentation task, the F1-max score of DCP-SFR dramatically increases from a suboptimal 44.9% in the 1-shot setting to a state-of-the-art 47.1% in the 4-shot setting. The results demonstrate that DCP-SFR is able to effectively leverage the reference data, leading to a more precise and reliable final segmentation.

4.3. Qualitative Comparison

We visualize the segmentation outputs of competing methods on representative samples, and Figure 3 clearly demonstrates the superior capability of DCP-SFR in handling diverse and subtle defect types. In the ‘Capsule’ case, either AnomalyGPT and PromptAD either produce diffuse, noisy heatmaps or fail to capture the full extent of the faint imprint defect. In contrast, DCP-SFR is able to generate a clean and precise result that closely matches the ground truth. Similarly, for the ‘Hazelnut’ example, WinCLIP+ and PromptAD

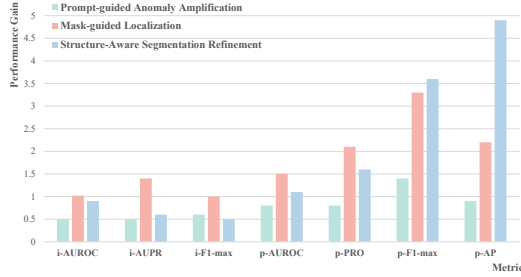


Figure 4. Performance gains from proposed modules.

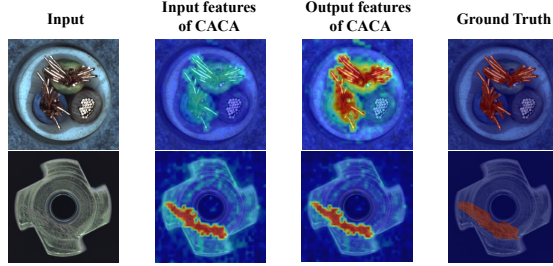


Figure 5. Visualization of CACA input and output features. The results show that CACA effectively amplifies of subtle anomalies (row 1) and suppresses of background noise (row 2).

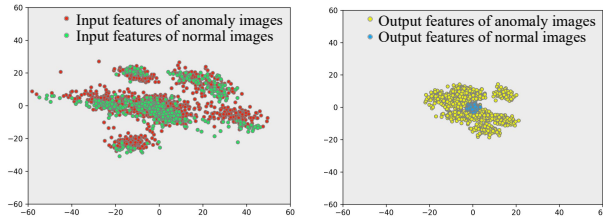


Figure 6. PCA visualization of ACL input and output features.

tAD tend to produce overly large and unfocused activation areas around the small hole, while our method provides a compact and accurately bounded heatmap. Finally, in the complex ‘Pill’ case with multiple small defects, DCP-SFR successfully identifies all distinct contaminated spots with sharp boundaries, whereas other methods often merge them into a single, blurry region or miss some of them. The results demonstrate the superior capability of DCP-SFR in producing precise segmentation maps with sharp boundaries, underscoring the effectiveness of our proposed defect cue-preservation and refinement strategies.

Table 2. Ablation study on the VisA under the 1-shot setting.

	VisA Classification			VisA Segmentation			
	i-AUROC	i-AUPR	i-F1-max	p-AUROC	p-PRO	p-F1-max	p-AP
DCP-SFR	92.0	93.7	87.8	98.2	93.3	44.6	39.9
w/o SSR	91.1	93.1	87.3	97.1	91.7	41.0	35.0
w/o ACL	89.7	92.3	86.8	96.7	91.2	41.3	37.7
w/o CACA	91.5	93.2	87.2	97.4	92.5	43.2	39.0

4.4. Ablation Study

To validate the contribution of core components in DCP-SFR, we conduct a series of ablation studies on the VisA dataset under the 1-shot setting. We create three variants

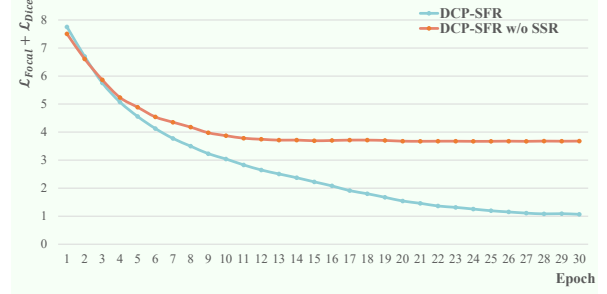


Figure 7. The impact of SSR on the training dynamic.

by disabling one module at a time: ‘DCP-SFR w/o CACA’, ‘DCP-SFR w/o ACL’ and ‘DCP-SFR w/o SSR’. The results are summarized in Table 2 and visualized in Figure 4.

Removing CACA leads to a significant degradation in segmentation performance. As shown in Table 2, the p-F1-max score drops sharply from 44.6% to 43.2%, indicating that without initial cue amplification, the model struggles with small or low-contrast anomalies, and this can be visually confirmed in Figure 5. The absence of ACL results in the most substantial performance drop, with the p-AP score decreasing from 39.9% to 37.7%. The results highlight the ACL’s crucial role in accurate anomaly localization and background suppression. The PCA projections in Figure 6 further reveals that while the pre-localization features of normal and anomaly samples are heavily mixed, the post-localization features show a distinct separation: normal features collapse into a tight cluster near the origin, while anomaly features remain scattered, evidencing the powerful capability of ACL in learning a discriminative feature space. Disabling SSR also causes a notable decline in segmentation performance, with the p-AUROC score falling from 98.2% to 97.1%. The impact of SSR on the training dynamic is illustrated in Figure 7, and we can observe that the full model only converges to a better solution but also exhibits a faster convergence speed.

5. Conclusion

We presented DCP-SFR, a few-shot anomaly detection framework to address the issue of defect cue fading in deep feature extraction. By incorporating a conditional amplification module, the defect cues are enhanced in the initial signals derived from learnable prompts. Built on the defect cues, a reconstruction-based localization module is employed to further distinguish anomalous regions from normal ones. By performing a structure-aware segmentation refinement, DCP-SFR outperforms state-of-the-art methods in the subtasks of image classification and anomaly segmentation. Extensive experiments demonstrate the model’s generalization capabilities across diverse object categories and anomaly types, confirming its potential as a scalable and flexible solution for diverse few-shot learning tasks.

Acknowledgement

This work was supported in part by TCL Science and Technology Innovation Fund (Project No. 20231752), in part by the National Key Research and Development Program of China (Grant No. 2024YFE0105400), in part by National Natural Science Foundation of China (Grants No. 62372186 and No. 62472179), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11206622), in part by the Guangdong Provincial Science and Technology Program (Project No. 2025A050508003 and No. 2025A0505020016), in part by the Guangdong Basic and Applied Basic Research Foundation (Project No. 2024A1515011437), in part by State Key Laboratory of Advanced Papermaking and Paper-based Materials (Project No. 2025PT09 and No. 2024ZD08).

References

- [1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 1
- [2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 6
- [3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 6
- [4] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, 2024. 1
- [5] 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 12: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv e-prints*, abs/2305.17382, 2023. 3, 6
- [6] 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. *arXiv e-prints*, abs/2311.00453, 2023. 1
- [7] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv e-prints*, abs/2005.02357, 2020. 3
- [8] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. 2, 3, 6
- [9] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition International*, 2021. 3
- [10] Chenghao Deng, Haote Xu, Xiaolu Chen, Haodi Xu, Xiaotong Tu, Xinghao Ding, and Yue Huang. Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection. In *Proceedings of ACM International Conference on Multimedia*, 2024. 3
- [11] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. *A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data*. 2002. 3
- [12] 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*, 2022. 1
- [13] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiguang Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2
- [14] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 2, 3, 6
- [15] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 1
- [16] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, and Yabiao Wang et al. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *AAAI Conference on Artificial Intelligence*, 2023. 1, 2
- [17] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael W. Spratling, and Yanfeng Wang. Registration based few-shot anomaly detection. *European Conference on Computer Vision*, 2022. 1, 2, 3
- [18] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 6
- [19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv e-prints*, abs/1412.6980, 2014. 7
- [20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 2023. 3
- [21] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022. 1

- [22] Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1
- [23] Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. In *Advances in Neural Information Processing Systems*, 2023. 1
- [24] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv e-prints*, abs/1911.02855, 2019. 6
- [25] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, and Yanyun Qu et al. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 6
- [26] Yiyue Li, Shaoting Zhang, Kang Li, and Qicheng Lao. One-to-normal: Anomaly personalization for few-shot anomaly detection. In *Advances in Neural Information Processing Systems*, 2024. 3
- [27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017. 6
- [28] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, and Klaus-Robert Müller et al. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *Transactions on Machine Learning Research*, 2022. 1
- [29] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenying Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3
- [30] 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 *International Conference on Learning Representations*, 2023. 1
- [31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, and Alaaeldin El-Nouby et al. Dinov2: Learning robust visual features without supervision. *arXiv e-prints*, abs/2304.07193, 2023. 2, 3
- [32] Jonathan Pirnay and Keng Yip Chai. Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing*, 2021. 1
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, and Gabriel Goh et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3
- [34] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Scholkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3
- [35] Han Sun, Yunkang Cao, Hao Dong, and Olga Fink. Anomaly anything: Promptable unseen visual anomaly generation. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2025. 1, 2
- [36] Ze Wang, Yipin Zhou, Rui Wang, Tsung-Yu Lin, Ashish Shah, and Ser-Nam Lim. Few-shot fast-adaptive anomaly detection. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022. 2
- [37] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 1
- [38] Guoyang Xie, Jingbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. In *The International Conference on Learning Representations*, 2023. 3
- [39] Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2021. 1
- [40] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1
- [41] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2022. 2
- [42] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem – a discriminatively trained reconstruction embedding for surface anomaly detection. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2
- [43] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 1
- [44] Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3
- [45] Bingke Zhu, Zhaopeng Gu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Adformer: Generalizable few-shot anomaly detection with dual cnn-transformer architecture. *IEEE Transactions on Instrumentation and Measurement*, 74:1–16, 2025. 3
- [46] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 6
- [47] Jiawen Zhu, Yew Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *IEEE/CVF International Conference on Computer Vision*, 2025. 1
- [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*, 2022. 1, 2, 6