electronics

MDPI

Article

BAG-CLIP: Bifurcated Attention Graph-Enhanced CLIP for Zero-Shot Industrial Anomaly Detection

Hua Wu, Tingting Zhang * and Shubo Li

School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China; wuhua@ncepu.edu.cn (H.W.); 120242227387@ncepu.edu.cn (S.L.)
* Correspondence: 120232227042@ncepu.edu.cn

Abstract

While vision-language models (VLMs) have been widely applied in zero-shot anomaly detection (ZSAD), their performance remains limited by the inability to distinguish fine-grained normal and abnormal textures, coupled with inadequate capabilities in detecting complex morphological anomalies. To address these limitations, this paper proposes BAG-CLIP (Bifurcated Attention Graph-Enhanced CLIP), a dual-path graph-enhanced zero-shot anomaly detection method. This approach employs a Bifurcated Self-Attention (BSA) module to decouple visual features, processing global semantics and spatial details separately to mitigate the inherent conflict between abstract semantic representation and precise spatial localization. A Self-Attention Graph (SAG) module is designed to model the topological structure of complex morphological anomalies. This module dynamically constructs visual features' topological relationships and utilizes graph convolutions to aggregate neighborhood information, thereby enhancing the model's representational capacity for diverse and complex morphological anomalies. Extensive experiments are conducted on five diverse industrial datasets, featuring complex transmission line backgrounds alongside general industrial scenarios. The proposed method is comprehensively evaluated against 11 state-of-the-art (SOTA) methods. On the EPED (Electrical Power Equipment Dataset) and MPDD datasets, BAG-CLIP outperforms the second-best methods in image-level AUROC (Area Under the Receiver Operating Characteristic Curve) by 3.7% and 2.8%, respectively. BAG-CLIP achieves superior performance in both zero-shot anomaly detection and segmentation.

Keywords: zero-shot anomaly detection; topological structure modeling; feature decoupling; vision-language model

Academic Editor: Ping-Feng Pai

Received: 16 March 2026
Revised: 6 April 2026
Accepted: 13 April 2026
Published: 15 April 2026

1. Introduction

Visual anomaly detection serves as a core component of industrial production and infrastructure maintenance, which is critical to ensuring production safety and the stable operation of the national economy [1]. Its core task is to identify and localize deviations from "normal" samples. However, traditional anomaly detection methods, whether unsupervised or semi-supervised, typically require training a dedicated model for each category to be inspected [2]. This approach necessitates a large number of normal samples for distribution modeling and frequently requires labeled abnormal samples, making it difficult to address the critical issue of diverse anomaly types in industrial applications [3]. In contrast, zero-shot anomaly detection enables anomaly detection without target category calibration

or training [4]. Vision-language Models (VLMs) typified by CLIP have become the mainstream paradigm for realizing zero-shot anomaly detection [5]. CLIP leverages rich and generalizable cross-modal knowledge to align visual features with text prompts. However, applying existing zero-shot anomaly detection methods to complex industrial scenarios still exhibits the following limitations: First, although CLIP encodes generalizable knowledge, it still suffers from feature misalignment in specific anomaly detection tasks. Second, while existing methods can implicitly capture certain spatial relationships through self-attention mechanisms or feature distributions, they still struggle to explicitly construct and leverage the dynamic topological connections among fine-grained normal and abnormal textures [6]. For industrial anomalies with complex and irregular morphologies, such as continuous fine cracks, winding scratches, or subtle structural deformations, this local topological information is crucial. Relying solely on global semantic understanding is insufficient to precisely localize these morphologically complex anomalies. This ultimately leads to a significant decline in pixel-level anomaly segmentation performance.

To address the aforementioned problems, this paper proposes a Bifurcated Attention Graph-Enhanced CLIP framework (BAG-CLIP). This framework aims to achieve more accurate detection of complex morphological industrial anomalies by enhancing the discriminative power and contextual awareness of visual features. Extensive experiments are conducted on multiple datasets, including the electrical power equipment inspection dataset EPED [7], the transmission line inspection dataset InsPLAD [8], the metal parts dataset MPDD [9], the general industrial product dataset BTAD [10], and the real-world industrial dataset MIAD [11]. The results demonstrate that BAG-CLIP significantly outperforms current state-of-the-art methods, such as WinCLIP [12], AnomalyCLIP [13], Bayes-PFL [14], and INP-Former [15], in both image-level anomaly classification and pixel-level anomaly segmentation tasks, validating the effectiveness and advancement of the proposed method.

The main contributions of this paper are as follows:

1. We introduce a Bifurcated Self-Attention (BSA) module integrated directly into the image encoder to explicitly decouple visual feature processing into two distinct pathways: a Global Semantic Branch and a Detail-Preserving Branch. The former is dedicated to extracting global contextual information, while the latter retains high-fidelity, fine-grained structural and textural information. This dual-path design effectively mitigates the intrinsic conflict between global semantic abstraction and spatial localization.
2. We propose a Self-Attention Graph (SAG) module to model complex morphological anomalies, such as fine cracks and winding scratches. By treating the spatial features extracted from the Detail-Preserving Branch as graph nodes, this module dynamically infers their topological relationships and aggregates neighborhood information. This mechanism effectively suppresses background interference while significantly enhancing the model's capacity to represent the topological structures of these complex anomalies.
3. We conduct extensive experiments on five diverse industrial datasets, notably including challenging transmission line inspection datasets. We perform comprehensive comparisons with 11 state-of-the-art zero-shot and few-shot methods. Experimental results validate that the proposed method delivers superior zero-shot anomaly detection (ZSAD) performance in both image-level anomaly classification and pixel-level anomaly segmentation.

2. Related Work

2.1. Traditional Anomaly Detection

To address the problem where normal samples are easily obtained while abnormal samples are scarce during quality control of specific product categories in industrial production, traditional anomaly detection methods emerged. These methods usually train a dedicated model for each category to be inspected and can be primarily divided into two categories: reconstruction-based and feature embedding-based methods [16]. Reconstruction-based methods, such as Autoencoders [17] and Generative Adversarial Networks (GANs) [18], often suffer from overly strong generalization capabilities. Consequently, they may reconstruct abnormal regions alongside normal ones, leading to missed detections that limit their application in high-precision industrial scenarios [19]. Feature embedding-based methods, such as Support Vector Data Description (SVDD) [20] and Deep SVDD [21], require training separate models for each category, rely on handcrafted features, and possess overly complex network structures, resulting in limited generalization capability and flexibility [22]. To address the scarcity of abnormal samples, few-shot anomaly detection methods, represented by PatchCore [23] and PaDiM [24], have established strong baselines by introducing a pre-trained feature memory bank and a multivariate Gaussian distribution modeling mechanism, respectively. However, the detection performance of these approaches still relies on utilizing normal samples from the target category to construct the feature distribution. Consequently, when applied to Unmanned Aerial Vehicle (UAV) inspection tasks where data for the target category is absent, they fail to achieve effective detection of unseen categories. Furthermore, neither of the aforementioned traditional methods can achieve effective zero-shot anomaly detection.

2.2. CLIP-Based Zero-Shot Anomaly Detection

In recent years, emerging zero-shot anomaly detection methods built on CLIP provide an effective solution to the above challenges, benefiting from their strong cross-modal semantic alignment ability [12,13]. In traditional computer vision, zero-shot generally denotes generalization to unseen object categories. Recent approaches extend this definition to generalization across entirely unseen datasets and specific downstream tasks. When applied to industrial anomaly detection, these methods can accurately detect and locate anomaly regions in novel-category samples without using any training data from the target domain. For instance, WinCLIP achieves anomaly detection by designing combined text prompts for “normal” and “abnormal” states and aggregating multi-scale visual features for matching. However, its performance is limited by the pre-training characteristics of underlying VLMs on natural images, and its reliance on handcrafted text prompts results in poor adaptability when faced with unknown anomaly categories. Furthermore, APRIL-GAN [25] and AnomalyCLIP address the zero-shot anomaly detection problem by fine-tuning CLIP on anomaly detection datasets containing anomaly labels. APRIL-GAN employs manually designed fixed text prompts, which reduces the detection capability for unknown object categories.

To solve the difficulty in designing prompt engineering, AdaCLIP [26] incorporates learnable static and dynamic prompts into CLIP; however, feature discriminability still requires improvement when dealing with diverse and complex anomalies in intricate industrial scenarios. AnomalyCLIP uses handcrafted prompts that require rich expert knowledge, and a single form of learnable prompt struggles to capture complex anomaly semantics, while an unconstrained prompt space limits the model’s generalization ability to unseen categories. PPTA-CLIP [27] enhances learnable text prompts by incorporating local visual patch information. DyC-CLIP [28] proposes a Frequency-domain Dynamic Adapter (FDA) to integrate global visual information into textual prompts. GenCLIP [29] performs

dual-branch inference by ensembling a vision-enhanced branch and a query-only branch. However, the query-only branch relies on a General Query Prompt (GQP), which leads to lower detection accuracy for unseen categories in power transmission line scenarios. Furthermore, while CLIP-AD [30] introduces a Staged Dual-Path (SDP) model to fuse multi-scale visual features, its fusion mechanism solely relies on residual superposition. This limitation tends to result in blurred edges of anomalous regions or missed detections of subtle anomalies. To address these issues, the Bayesian Prompt Flow Learning (Bayes-PFL) method models the prompt space as a learnable probability distribution, but this comes with increased computational complexity [14]. Although it shows significant improvement in sample dependency, missed detections still occur when detecting "anomalies highly similar to normal background features," where abnormal regions are not effectively identified. AdaptCLIP [31] treats the CLIP model as a foundational service, incorporating lightweight visual, textual, and prompt-query adapters to alternately learn adaptive representations. AF-CLIP [32] enhances visual representations by introducing a lightweight adapter that simultaneously optimizes class-level and patch-level features alongside learnable textual prompts. Tipsomaly [33] addresses spatial misalignment by leveraging TIPS, which is a vision-language model trained with spatially aware objectives, and employs decoupled prompts to bridge the global-local feature distribution gap via local evidence injection into global anomaly scores. UniADet [34] decouples classification and segmentation across multi-scale hierarchical features to improve anomaly detection performance.

Despite significant progress in prompt learning and sample dependency by existing CLIP-based zero-shot anomaly detection methods, the text encoders they rely on struggle to distinguish fine-grained texture differences and ignore the spatial structural correlations between anomalies and their surrounding areas. Consequently, their detection performance drops significantly when confronted with subtle anomalies similar to background textures or anomalies with complex morphologies, such as cracks, missing parts, wear, rust, and scratches. Therefore, the BAG-CLIP framework proposed in this paper aims to solve the above problems by introducing the BSA module to decouple global semantics from spatial details and designing the SAG module to model the topological structural relationships, thereby significantly improving the model's representation and detection capabilities for complex morphological anomalies.

3. Proposed Method

3.1. Overall Architecture

The overall architecture of BAG-CLIP is illustrated in Figure 1. This framework is designed to address the challenges of zero-shot anomaly detection by explicitly decoupling semantic and fine-grained detail features, followed by modeling the topological contextual relationships among these detail features.

The detailed processing pipeline of BAG-CLIP proceeds as follows. Initially, an input inspection image is passed through a pre-trained CLIP image encoder to extract multi-level visual features. These multi-level features are then fed into the BSA module. Activated in the deeper layers of the Vision Transformer (ViT), the BSA module bifurcates the feature extraction process into two functionally independent paths. This design explicitly mitigates the inherent conflict between high-level semantic abstraction and precise spatial localization.

- Global Semantic Branch: This path is responsible for aggregating high-level contextual information to generate a compact feature vector for image-level classification.
- Detail-Preserving Branch: This path preserves and concatenates multi-scale, high-resolution feature maps to generate local detail features x_{detail} for pixel-level anomaly localization.

While x_{detail} preserves abundant spatial information, it lacks the context necessary to accurately characterize complex defect morphologies. Therefore, we design the SAG module. As shown in Figure 2, the SAG module conceptualizes the N image patch features within x_{detail} as graph nodes. It dynamically constructs their adjacency relationships and performs information propagation via a multi-layer Graph Convolutional Network (GCN). This mechanism models and aggregates the contextual associations and structural correlations among the features, ultimately outputting the enhanced local features $\boldsymbol{F}_{SAG}$.

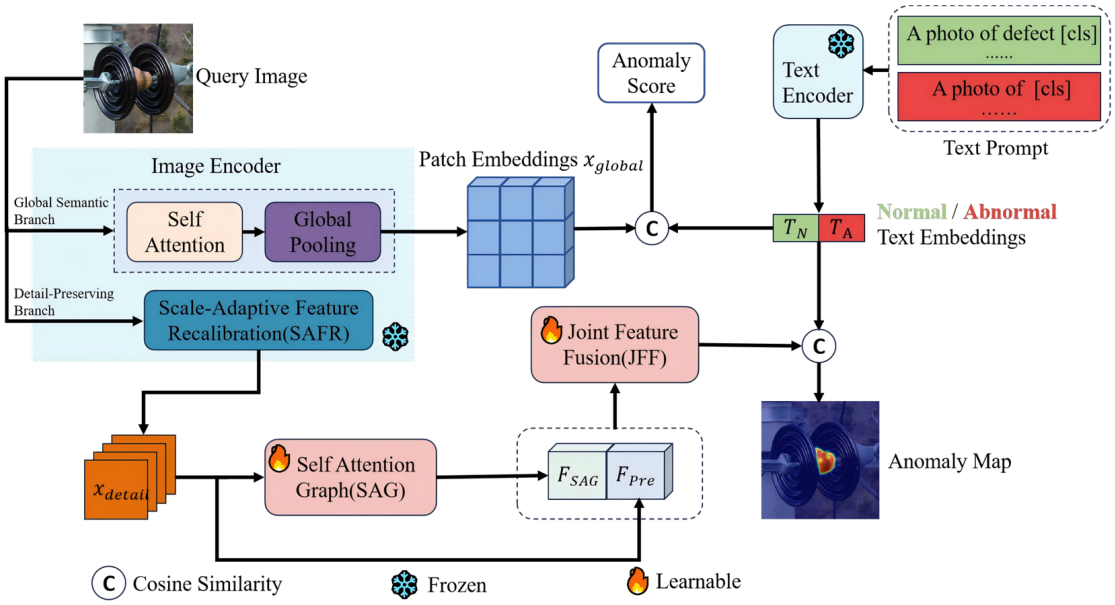

Figure 1. The architecture of BAG-CLIP.

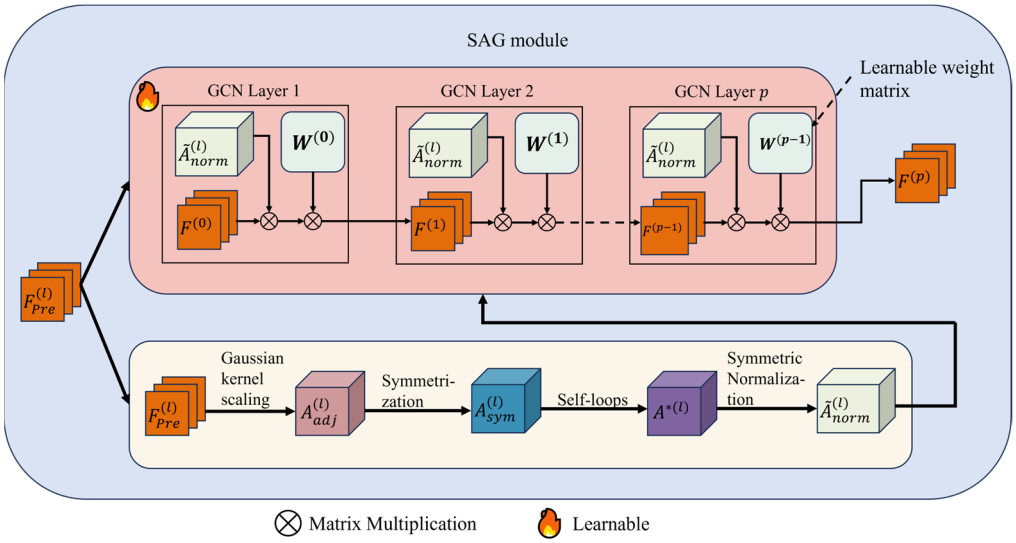

Figure 2. Overview of the SAG module.

Next, to integrate the original knowledge from the pre-trained model with the topological context mined by the SAG module, we introduce the Joint Feature Fusion (JFF) module. The JFF module concatenates the original local features with the enhanced features, ensuring that the resultant representation encapsulates both high-fidelity visual details and structured contextual awareness. Subsequently, a trainable projection layer maps these fused features into a joint semantic space aligned with the CLIP text embeddings, yielding the final visual representation, $\boldsymbol{\Psi}_{final}$.

Concurrently, the CLIP text encoder processes state-specific prompt templates to generate corresponding text features $(\boldsymbol{T}_{\mathrm{N}}, \boldsymbol{T}_{\mathrm{A}})$ representing the "normal" and "abnormal" conditions, respectively. Following WinCLIP, we decouple the prompts into image-level contextual templates and semantic state descriptions. The normal descriptions are formulated using adjectives such as "flawless [cls]", "[cls] without defect", "unblemished [cls]" and "perfect [cls]". Conversely, the abnormal descriptions comprise descriptors like "damaged [cls]", "[cls] with defect", "[cls] with flaw" and "broken [cls]". In these templates, "[cls]" serves as the specific class name of the target object being evaluated (e.g., "insulator", "fender hammer"). These state descriptions are subsequently integrated into contextual templates to form the final comprehensive text prompts, such as "a photo of defect [cls]" and "a photo of a flawless [cls]". Subsequently, the text embeddings extracted from all combinations corresponding to the same state are aggregated via mean pooling and followed by L2 normalization to formulate the final robust text features $(\boldsymbol{T}_{\mathrm{N}}, \boldsymbol{T}_{\mathrm{A}})$.

Finally, by calculating the cosine similarity between $\boldsymbol{\Psi}_{final}$ and these text features $(\boldsymbol{T}_{\mathrm{N}}, \boldsymbol{T}_{\mathrm{A}})$, high-resolution anomaly maps are generated, allowing precise zero-shot anomaly detection.

3.2. Bifurcated Self-Attention (BSA)

The BSA module is activated in the final Transformer layers of the image encoder. Its core function is to explicitly divide feature processing into two complementary and independent branches: the Detail-Preserving Branch and the Global Semantic Branch.

3.2.1. Detail-Preserving Branch

Detail-Preserving Branch is designed to maximally preserve fine-grained structural and textural information at high resolution, which is essential for subsequent pixel-level anomaly localization and dynamic graph construction. We extract feature maps $\boldsymbol{\Phi}_k$ from K different intermediate layers of the ViT encoder, where $k = 1, \ldots, K$. These multi-level feature maps are concatenated to form a comprehensive multi-scale local feature representation $x_{concat} \in \mathbb{R}^{C \times H \times W}$. However, direct cross-layer feature integration may introduce scale redundancy and semantic misalignment. We introduce the Scale-Adaptive Feature Recalibration (SAFR) module to dynamically reweight channels, suppressing redundant noise and highlighting anomaly-relevant features.

Specifically, the SAFR module employs Global Average Pooling (GAP) to compress the spatial dimensions of x_{concat}, generating a channel-wise statistic descriptor. Grounded in the principle of global information embedding [35], this design addresses the limitation that local patch features primarily encode regional information and may lack global context. Therefore, GAP condenses the global spatial information into a unified channel descriptor. These channel-wise statistics aggregate local descriptors into a representation capturing overall image texture. For anomaly detection, where anomalies appear at arbitrary locations, this operation mitigates spatial variance and introduces translation invariance, directing the subsequent MLP to focus primarily on inter-channel dependencies. This enables the model to adaptively highlight anomaly-sensitive texture channels, without relying on spatial coordinates. This descriptor is subsequently fed into a bottleneck Multi-Layer Perceptron (MLP) with a dimensionality reduction ratio r, followed by a Sigmoid activation function to learn the adaptive channel weights. The recalibrated detail-preserving feature x_{detail} is formulated as:

$$x_{detail} = x_{concat} \otimes \sigma(W_2 \delta(W_1 \cdot GAP(x_{concat}))) \tag{1}$$

where δ denotes the ReLU activation function, σ represents the Sigmoid function, while $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are the learnable weight matrices of the MLP. The dimensional-

ity reduction ratio r is empirically set to 16, in order to achieve a favorable trade-off between channel-wise feature recalibration performance and computational complexity [36].

Through the SAFR module, the model adaptively purifies the multi-scale representations, providing a robust and high-fidelity spatial feature foundation for the subsequent SAG module.

3.2.2. Global Semantic Branch

The Global Semantic Branch focuses on aggregating global contextual information to construct a high-level semantic understanding of the entire image, which primarily serves the image-level anomaly classification task. To achieve the sequential aggregation of global semantics, we employ a standard Transformer block to process the input features.

First, the feature passes through a Multi-Head Self-Attention module equipped with a residual connection, yielding an intermediate representation x'_{global}. Subsequently, this intermediate feature is processed by an MLP block (incorporating Layer Normalization) to produce the final global semantic output of the i-th layer, denoted as x^i_{global}. This process ensures an effective refinement of the high-level semantics. The computations are as follows:

$$x'_{global} = x^{i-1}_{global} + softmax(\frac{QK^T}{\sqrt{d_v}})V \tag{2}$$

$$x^i_{global} = x'_{global} + MLP(LN(x'_{global})) \tag{3}$$

Through the deliberate design of the BSA module, we achieve an effective decoupling of spatial details x_{detail} and global semantics x_{global}, overcoming the limitation of a single feature stream struggling to balance both tasks. This provides the most suitable feature support for pixel-level localization and image-level classification, respectively.

3.3. Self-Attention Graph (SAG)

To model the non-Euclidean structure among the local features x_{detail}, we design the SAG module. The core idea of the SAG module is to treat the local features x_{detail} as the set of nodes $\mathcal{V}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Let $\boldsymbol{V}^{(0)}$ be the $N \times D$ dimensional feature matrix corresponding to x_{detail}, where N is the number of image patches and D is the dimension of the feature. $v_i^{(0)}$ represents the initial feature of the i-th node.

The connectivity of the graph $\mathcal{E}$ should be dynamically determined by the similarity of the features. Image patches that are close in the feature space, regardless of their physical location in the image, are more likely to belong to the same semantic region. This dynamic construction, based on feature similarity rather than fixed spatial adjacency, can more accurately capture the non-local correlations between complex-morphology anomalies and normal regions.

Let $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N]^\top \in \mathbb{R}^{N \times d}$ denote the feature matrix extracted from the intermediate layer of the image encoder. To characterize the manifold structure and mitigate the curse of dimensionality, we employ a learnable metric parameterized by a projection matrix $\boldsymbol{W}_\Phi \in \mathbb{R}^{d \times d'}$. The semantic affinity $\mathcal{S}_{ij}$ between nodes v_i and v_j is formulated using a Generalized Gaussian Kernel with adaptive local scaling, which normalizes the density variations in the feature space:

$$\mathcal{S}_{ij} = \exp\left(-\frac{\|\boldsymbol{W}_\Phi \boldsymbol{h}_i - \boldsymbol{W}_\Phi \boldsymbol{h}_j\|_2^2}{\sigma_i \sigma_j}\right) \tag{4}$$

where σ_i and σ_j denote the local scale parameters for v_i and v_j, respectively, each often defined as the distance to its K-th nearest neighbor. This adaptive kernel ensures that topological connections are robust to the varying texture densities inherent in industrial surfaces.

To induce a sparse graph topology that preserves local manifold structures while filtering noise, we dynamically construct the edge set using a soft thresholding function based on the computed affinity. The threshold is determined dynamically for each node using the K-Nearest Neighbors (KNN) algorithm. Specifically, we set a node-specific threshold τ_i defined as the semantic affinity between node v_i and its K-th nearest neighbor. Based on this, the initial adjacency matrix $\boldsymbol{A}_{adj}$ is constructed:

$$(\boldsymbol{A}_{adj})_{ij} = \mathrm{ReLU}(S_{ij} - \tau_i) \tag{5}$$

However, this initial adjacency matrix $\boldsymbol{A}_{adj}$ is directed and lacks self-connections, making it unsuitable for direct use in standard GCNs. We therefore refine the graph structure through the following three logical steps:

- Symmetrization: To ensure information can flow bidirectionally between neighbors, the adjacency matrix is converted to a symmetric form.

$$\boldsymbol{A}_{sym} = \boldsymbol{A}_{adj} + \boldsymbol{A}_{adj}^{\top} \tag{6}$$

Creating an undirected graph: This aligns with the requirement of standard GCNs for fair information exchange.

- Introduction of Self-loops: To ensure each node retains its own original features during information aggregation, preventing its own information from being overly diluted during multi-layer propagation, we add the identity matrix $\boldsymbol{I}_N$:

$$\boldsymbol{A}^* = \boldsymbol{A}_{sym} + \boldsymbol{I}_N \tag{7}$$

- Symmetric Normalization: To prevent gradient vanishing or explosion during multi-layer GCN propagation, which affects model training stability, we employ symmetric normalization. First, we compute the degree matrix $\boldsymbol{D}^*$, whose diagonal elements $(\boldsymbol{D}^*)_{ii}$ are defined as the sum of the i-th row of $\boldsymbol{A}^*$. Based on the degree matrix $\boldsymbol{D}^*$, we compute the final normalized adjacency matrix $\widetilde{\boldsymbol{A}}_{norm}$ used for GCN propagation:

$$(\boldsymbol{D}^*)_{ii} = \sum_{j=1}^{N} (\boldsymbol{A}^*)_{ij} \tag{8}$$

$$\widetilde{\boldsymbol{A}}_{norm} = (\boldsymbol{D}^*)^{-1/2} \boldsymbol{A}^* (\boldsymbol{D}^*)^{-1/2} \tag{9}$$

Through these steps, we construct a dynamic graph structure that effectively captures feature space locality and is suitable for stable information propagation.

3.4. Graph-Based Information Propagation

Using the dynamically constructed graph structure $\widetilde{\boldsymbol{A}}_{norm}$, we employ a P-layer Graph Convolutional Network to iteratively update and enhance the node features. The input to the GCN is $\boldsymbol{F}^{(0)}$ (the feature matrix of x_{detail}). The feature propagation rule from layer p to layer $p+1$ is defined as:

$$\boldsymbol{F}^{(p+1)} = \mathrm{ReLU}(\widetilde{\boldsymbol{A}}_{norm} \boldsymbol{F}^{(p)} \boldsymbol{W}^{(p)}) \tag{10}$$

where $\boldsymbol{F}^{(p)}$ is the node feature matrix at layer p, $\boldsymbol{W}^{(p)}$ is the learnable weight matrix for that layer, and ReLU is the non-linear activation function. By stacking p convolutional layers, each node feature can iteratively aggregate information from its p-hop neighborhood,

thereby capturing a wider context. The final output of the SAG module is the enhanced local feature matrix $\boldsymbol{F}_{SAG} = \boldsymbol{F}^{(P)}$. Through graph information propagation, this feature deeply mines the associations between anomalies and their surrounding regions, effectively learning the topological structure among spatial details.

3.5. Joint Feature Fusion (JFF)

To combine the pre-trained knowledge in the original image features with the topological structure information in the enhanced features, we propose the Joint Feature Fusion (JFF) module. As shown in Figure 1, the original local features $\boldsymbol{F}_{Pre}$ (the feature matrix of x_{detail}) contain rich foundational visual information, while the enhanced features $\boldsymbol{F}_{SAG}$ have aggregated local topological context via graph convolution, providing a more precise representation of complex-morphology anomalies. To improve the model's adaptability to diverse anomaly morphologies, we avoid fixed scalar coefficients for fusion ratio control. Instead, we concatenate the original and enhanced features, followed by a learnable linear projection via a fully connected layer. This design eliminates manual hyperparameter tuning, allowing the model to automatically learn optimal fusion weights for the two feature pathways across different dimensions via backpropagation [37,38].

We first fuse the two types of features using feature concatenation, creating a higher-dimensional, more informative merged feature $\boldsymbol{F}_{merged}$:

$$\boldsymbol{F}_{merged}^{(l)} = Concat(F_{Pre}^{(l)}, F_{SAG}^{(l)}) \in \mathbb{R}^{N\times(C_P+C_S)} \tag{11}$$

Subsequently, the merged high-dimensional feature $\boldsymbol{F}_{merged}$ is passed through a trainable linear projection layer (with weights $\boldsymbol{W}_{proj}$). This projection layer not only unifies the feature dimensions but, more critically, maps the jointly fused features into the joint semantic space aligned with the text features. Finally, the projected features undergo L2 normalization to ensure the accuracy and stability of the subsequent cosine similarity calculations. The computation for the final fused and projected feature Ψ_{final} is as follows:

$$\boldsymbol{\Psi}_{final} = \frac{\boldsymbol{F}_{merged}\boldsymbol{W}_{proj}}{\|\boldsymbol{F}_{merged}\boldsymbol{W}_{proj}\|_2} \tag{12}$$

where $\|\cdot\|_2$ denotes the L2 norm.

3.6. Loss Function

To effectively train the BAG-CLIP framework to handle the dual tasks of image-level anomaly classification and pixel-level anomaly segmentation, we designed a composite objective function $\mathcal{L}_{obj}$. This function consists of two components: a loss $\mathcal{L}_{CLS}$ for supervising image-level classification and a loss $\mathcal{L}_{SEG}$ for optimizing pixel-level segmentation. The total loss function is formulated as:

$$\mathcal{L}_{obj} = \mathcal{L}_{CLS} + \lambda\mathcal{L}_{SEG} \tag{13}$$

where λ represents the loss weighting between $\mathcal{L}_{CLS}$ and $\mathcal{L}_{SEG}$. The hyperparameter λ is set to 1 to balance the learning of global semantic classification and local detail segmentation. This ensures that the model effectively captures subtle local anomalies for pixel-level localization, while preventing it from overfitting to dataset-specific biases caused by excessive task-specific fine-tuning [5].

Considering that in anomaly detection datasets, normal samples often far outnumber abnormal samples, we use Focal Loss [39] to compute $\mathcal{L}_{CLS}$ to mitigate the class imbalance problem. Its formula is:

$$\mathcal{L}_{CLS} = -(1 - p_t)^{\beta} \log(p_t) \tag{14}$$

where p_t is the model's predicted probability for the correct class, and β is the focusing parameter.

The pixel-level anomaly segmentation loss $\mathcal{L}_{SEG}$ must address both pixel imbalance and region matching challenges. It is composed of a pixel-wise Focal Loss and a Dice Loss [40]. The Dice Loss optimizes segmentation accuracy by maximizing the Intersection over Union (IoU) between the prediction and the ground truth mask. The formula of $\mathcal{L}_{SEG}$ is:

$$\mathcal{L}_{SEG} = \mathcal{L}_{Focal}(\boldsymbol{S}, \boldsymbol{M}_{gt}) + \mathcal{L}_{Dice}(\boldsymbol{S}_{ab}, \boldsymbol{M}_{gt}) + \mathcal{L}_{Dice}(\boldsymbol{S}_{norm}, 1 - \boldsymbol{M}_{gt}) \tag{15}$$

where $\boldsymbol{S}$ is the model's output two-channel pixel-level prediction map, $\boldsymbol{M}_{gt}$ is the ground truth anomaly mask, and $\boldsymbol{S}_{ab}$ and $\boldsymbol{S}_{norm}$ represent the predicted probability maps for the abnormal and normal channels, respectively. By applying the Dice Loss constraint to both the abnormal and normal channels simultaneously, the model not only accurately localizes anomaly regions but also clearly defines the background, thereby generating more precise segmentation results.

4. Experiment and Analysis

4.1. Experimental Setup

4.1.1. Dataset Descriptions

To comprehensively and objectively evaluate the anomaly detection performance of the proposed model in power inspection and industrial scenarios, experiments were conducted using seven datasets: MVTec AD [41], VisA [42], EPED, InsPLAD, MPDD, BTAD and MIAD [11]. Among these, EPED and InsPLAD are datasets featuring power equipment with severe anomalies, covering categories such as clamps, vibration dampers, triangular plates, and insulators. These datasets contain images with diverse complex backgrounds, including grasslands, snowfields, forests, and transmission towers, as well as varying lighting conditions and shooting angles. MPDD is a metal parts anomaly detection dataset focused on anomalies in metal components; it includes common anomaly types such as scratches, cracks, dents, and abrasions. The images feature unified resolutions but diverse anomaly morphologies, allowing for the verification of the model's detection capabilities on metal surfaces. The BTAD dataset is a novel image dataset oriented towards real-world industrial scenarios. MIAD is a comprehensive dataset focused on outdoor maintenance inspection, featuring both surface and logical anomalies under complex, uncontrolled environmental conditions. The MVTec AD and VisA datasets contain anomaly sample annotations for various industrial products. In our experiments, they were used to train the SAG module and JFF module for the zero-shot anomaly detection task.

4.1.2. Evaluation Metrics

To effectively evaluate the anomaly localization and segmentation capabilities of the proposed model, AUROC (Area Under the Receiver Operating Characteristic Curve), F1-Max (maximum F1 score), AP (Average Precision) and AUPRO (Area Under Per-Region Overlap) were employed as evaluation metrics. AUROC reflects the classification performance of the model under different decision thresholds, with a value range of [0, 1]; a value closer to 1 indicates better classification performance, comprehensively measuring the model's sensitivity and specificity. F1-Max represents the maximum value of the F1 score across different decision thresholds, effectively balancing the model's precision and recall, making it suitable for class-imbalanced data scenarios. AP evaluates the model's ability to identify positive samples by calculating the area under the Precision-Recall curve, with higher values indicating higher identification accuracy for anomalous images. AUPRO

evaluates localization performance by measuring the area under the per-region overlap curve over the false positive rate range [0, 0.3]. By averaging the overlap over connected anomaly regions, it mitigates the bias toward large defect regions and provides a more balanced assessment of anomalies with different sizes.

4.1.3. Implementation Details

The experiments adopted the pre-trained CLIP (ViT-L/14@336px) as the default backbone, extracting feature embeddings from the 6th, 12th, 18th, and 24th layers. All images were resized to a resolution of 518 × 518 for both training and testing. For the zero-shot anomaly detection task, the model was trained on the MVTec AD and VisA datasets, and its performance was tested and verified on the EPED, InsPLAD, MPDD, BTAD and MIAD datasets. Rigorous verification confirms that there is no category overlap or intersection between the training and test sets, ensuring the fairness and reliability of the evaluation results. BAG-CLIP was trained for 5 epochs with a learning rate of 0.001. For the SAG module, the number of neighbors K is set to 15 and the number of GCN layers P is set to 5 by default. All experiments were conducted using PyTorch 1.9.2 (Meta Platforms, Menlo Park, CA, USA) on a single NVIDIA A6000 48GB GPU (NVIDIA, Santa Clara, CA, USA).

4.2. Comparative Experiments

To comprehensively and effectively evaluate the performance of BAG-CLIP in zero-shot anomaly detection tasks, it was compared with nine current state-of-the-art zero-shot anomaly detection methods, including WinCLIP, AnomalyCLIP, AdaCLIP, APRIL-GAN, INP-Former, Bayes-PFL, Tipsomaly, AdaptCLIP and AF-CLIP. To address the challenge of limited sample acquisition in power transmission line inspections, we introduce the strong few-shot anomaly detection baselines PatchCore and PaDiM for comparison. Considering the issue of sample scarcity in industrial scenarios, both methods are trained utilizing four normal samples. Experiments were conducted on five test datasets: EPED, InsPLAD, MPDD, BTAD and MIAD with performance evaluated from both image-level and pixel-level dimensions. The experimental results, as shown in Tables 1 and 2, demonstrate that BAG-CLIP exhibits significant performance advantages in both image-level classification and pixel-level segmentation tasks.

Table 1. Comparison of image-level results in different anomaly detection methods (AUROC, F1-Max, AP).

Method	EPED	InsPLAD	MPDD	BTAD	MIAD
WinCLIP	(59.9, 66.5, 54.2)	(78.3, 68.5, 65.4)	(61.5, 77.5, 69.2)	(68.2, 67.8, 70.9)	(61.3, 67.1, 61.8)
AnomalyCLIP	(61.6, 68.2, 62.4)	(68.6, 58.5, 55.5)	(77.5, 80.4, 82.5)	(88.2, 83.8, 88.2)	(63.5, 71.1, 64.4)
AdaCLIP	(69.3, 69.4, 70.2)	(85.6, 75.6, 71.9)	(73.0, 80.6, 76.3)	(89.2, 84.6, 89.8)	(65.8, 70.1, 66.5)
APRIL-GAN	(64.8, 67.7, 66.4)	(85.7, 72.6, 72.8)	(76.8, 80.7, 82.5)	(73.5, 68.0, 69.7)	(62.0, 68.9, 62.5)
INP-Former	(54.4, 64.8, 54.6)	(61.3, 56.6, 44.9)	(54.9, 74.6, 65.3)	(83.7, 84.6, 87.3)	(55.4, 67.2, 54.5)
Bayes-PFL	(67.8, 71.9, 64.0)	(80.4, 70.3, 64.0)	(76.9, 81.3, 78.7)	(90.5, 85.4, 89.1)	(60.5, 69.5, 60.8)
Tipsomaly	(72.7, 74.2, 71.7)	(70.0, 59.6, 59.3)	(75.5, 79.7, 79.8)	(94.9, 91.4, 95.5)	(73.6, 71.2, 71.9)
AdaptCLIP	(59.8, 67.0, 60.3)	(67.3, 61.8, 49.0)	(76.8, 79.3, 76.9)	(91.4, 90.3, 92.2)	(62.1, 71.0, 62.7)
AF-CLIP	(67.4, 71.1, 68.4)	(80.9, 75.2, 75.8)	(75.8, 81.6, 81.6)	(94.3, 91, 95.2)	(63.6, 70.4, 64.4)
PaDiM [4+]	(48.7, 64.5, 49.6)	(56.6, 55.0, 39.0)	(50.0, 74.0, 60.5)	(91.5, 86.7, 90.1)	(46.1, 65.9, 44.3)
PatchCore [4+]	(43.7, 63.7, 45.7)	(61.0, 59.4, 45.7)	(62.2, 79.9, 65.1)	(90.9, 88.2, 92.6)	(45.8, 60.4, 41.2)
BAG-CLIP (Ours)	(76.4 ± 0.2, 77.6 ± 0.1, 72.8 ± 0.3)	(88.4 ± 0.3, 75.6 ± 0.1, 75.9 ± 0.4)	(80.3 ± 0.3, 85.7 ± 0.1, 83.6 ± 0.1)	(92.2 ± 0.1, 92.7 ± 0.2, 96.9 ± 0.3)	(75.8 ± 0.3, 74.2 ± 0.4, 73.5 ± 0.1)

Note: Red indicates the best result, and Blue indicates the second-best result. [4+] denotes training with four normal samples. Results for BAG-CLIP are presented as mean ± standard deviation across five independent runs.

Table 2. Comparison of pixel-level results in different anomaly detection methods (AUROC, F1-Max, AP, AUPRO).

Method	EPED	InsPLAD	MPDD	BTAD	MIAD
WinCLIP	(62.9, 3.1, 3.8, 26.0)	(75.9, 20.9, 15.0, 39.7)	(71.2, 15.4, 14.1, 40.5)	(72.7, 18.5, 12.9, 27.5)	(67.4, 4.6, 3.9, 34.8)
AnomalyCLIP	(91.3, 21.5, 12.9, 74.7)	(85.4, 26.8, 20.1, 57.4)	(87.7, 30.6, 25.0, 73.3)	(87.7, 41.7, 38.5, 62.5)	(88.3, 9.0, 4.4, 71.5)
AdaCLIP	(94.8, 28.4, 24.4, 68.8)	(83.2, 23.3, 17.8, 53.6)	(93.9, 28.9, 25.8, 62.8)	(90.2, 40.6, 34.8, 20.3)	(80.6, 5.0, 1.6, 69.6)
APRIL-GAN	(94.9, 24.5, 18.9, 82.7)	(76.8, 16.3, 10.4, 43.8)	(94.3, 31.3, 26.6, 83.8)	(89.3, 40.6, 36.5, 68.8)	(87.7, 9.2, 4.2, 73.9)
INP-Former	(76.9, 2.2, 1.0, 40.5)	(82.4, 16.4, 10.2, 46.9)	(90.6, 14.4, 8.2, 71.1)	(88.6, 32.6, 26.7, 65.0)	(92.0, 8.5, 5.3, 42.4)
Bayes-PFL	(97.3, 22.6, 17.6, 67.7)	(91.4, 27.5, 21.1, 51.8)	(97.1, 32.9, 30.0, 84.6)	(92.3, 43.9, 40.2, 67.0)	(91.6, 6.6, 3.2, 65.2)
Tipsomaly	(93.5, 25.6, /, 81.2)	(86.9, 22.8, /, 55.1)	(95.7, 33.1, /, 86.5)	(96.7, 56.2, /, 84.7)	(91.3, 9.1, /, 70.5)
AdaptCLIP	(90.8, 26.4, 21.1, 50.5)	(83.7, 22.1, 15.7, 68.3)	(96.0, 30.4, 29.2, 92.9)	(94.8, 47.7, 44.8, 81.4)	(90.3, 7.6, 3.4, 69.7)
AF-CLIP	(96.1, 23.5, 19.7, 84.5)	(88.7, 27.2, 22.5, 61.8)	(96.6, 29.5, 27.9, 90.2)	(94.4, 47.5, 41.9, 78.4)	(91.6, 5.8, 2.7, 68.5)
PaDiM [4+]	(82.4, 5.0, 2.1, 50.0)	(78.7, 14.4, 8.6, 45.1)	(90.1, 15.9, 9.7, 65.7)	(96.3, 44.2, 38.9, 86.6)	(79.4, 3.7, 1.5, 58.2)
PatchCore [4+]	(75.3, 5.3, 2.0, 64.8)	(77.4, 16.0, 11.5, 54.2)	(93.8, 22.4, 19.7, 77.3)	(95.6, 45.1, 40.1, 80.5)	(72.9, 4.3, 2.1, 62.7)
BAG-CLIP (Ours)	(97.9 ± 0.2, 28.9 ± 0.2, 23.2 ± 0.3, 85.9 ± 0.4)	(92.5 ± 0.4, 29.3 ± 0.2, 22.7 ± 0.2, 62.6 ± 0.3)	(97.3 ± 0.1, 33.4 ± 0.2, 31.8 ± 0.1, 91.4 ± 0.3)	(96.5 ± 0.2, 49.6 ± 0.1, 45.7 ± 0.1, 80.7 ± 0.2)	(92.7 ± 0.4, 9.7 ± 0.2, 6.6 ± 0.2, 71.3 ± 0.3)

Note: Red indicates the best result, and Blue indicates the second-best result. [4+] denotes training with four normal samples. Results for BAG-CLIP are presented as mean ± standard deviation across five independent runs.

In the image-level anomaly classification task, BAG-CLIP achieved highly competitive and predominantly superior performance across all datasets. The EPED dataset contains a large number of power equipment images with complex morphological anomalies against backgrounds such as vegetation and snow. On the EPED dataset, BAG-CLIP achieved an image-level AUROC and F1-Max of 76.4% and 77.6%, respectively, representing improvements of 3.7% and 3.4% over the second-best method Tipsomaly. On the InsPLAD dataset, BAG-CLIP achieved an image-level AUROC of 88.4%. Compared with the few-shot anomaly detection approaches PatchCore and PaDiM, our method achieves AUROC improvements of 27.4% and 31.8%, respectively. This indicates that few-shot methods based on feature memory banks are overly dependent on the consistency of background distributions. Consequently, they are prone to failure caused by distribution estimation bias when confronted with complex and highly variable backgrounds in open-world scenarios. On the general industrial product dataset BTAD, BAG-CLIP also led with an F1-Max of 92.7% and an AP of 96.9%. On the MIAD dataset, BAG-CLIP achieves an AUROC of 75.8% and an AP of 73.5%, representing improvements of 13.7% and 10.8% over AdaptCLIP, respectively. The performance difference stems mainly from structural differences between the two frameworks. AdaptCLIP relies on lightweight bottleneck adapters that might attenuate fine-grained structural details under complex environments. In contrast, BAG-CLIP leverages a dual-path attention mechanism to better preserve detailed spatial context, providing stronger support for the detection of surface anomalies. Furthermore, the SAG module is incorporated to dynamically capture long-range semantic correlations, which facilitates the modeling and recognition of logical anomalies. These merits demonstrate the robust generalization ability of BAG-CLIP across diverse industrial scenarios.

In the pixel-level anomaly segmentation tasks, BAG-CLIP exhibits an equally significant advantage. Table 2 shows the anomaly detection results of BAG-CLIP with eleven competing methods over five industrial datasets. On the MIAD dataset, BAG-CLIP achieves top performance across three metrics, recording an AUROC of 92.7%. Regarding the power equipment InsPLAD dataset, the model achieves the highest AUROC of 92.5% and an F1-Max of 29.3%, representing improvements of 5.6% and 6.5% over Tipsomaly, respectively. For the MPDD dataset, BAG-CLIP maintains strong segmentation capability, with its AUROC reaching 97.3%. Similarly, on the EPED dataset, BAG-CLIP yields an AUROC of 97.9% and an F1-Max of 28.9%. These results reflect a 0.6% gain in AUROC over Bayes-PFL and a 3.3% gain in F1-Max over Tipsomaly. These consistent improvements indicate that the SAG module effectively utilizes fine-grained features from the detail-preserving branch to construct topological relations. By employing graph convolutional networks

to enhance visual representations, the proposed architecture provides better support for precise anomaly localization.

BAG-CLIP also significantly outperforms other methods in the localization of minute anomalies. As seen in rows 3 and 8 of Figure 3, WinCLIP and INP-Former exhibit numerous false positive regions in the target equipment and background when detecting power equipment anomalies with complex backgrounds, such as insulators and triangular plates; they fail to precisely localize anomalies, which is consistent with their performance on pixel-level metrics. As shown in row 4 of Figure 3, the heatmaps of APRIL-GAN appear significantly fragmented and noisy, with anomaly responses manifesting as scattered dots or small patches that fail to form a coherent region consistent with the true anomaly shape. This performance degradation is attributed to APRIL-GAN's use of fixed text prompts, which reduces its detection capability for unknown object categories and its resistance to background interference. As shown in row 6 of Figure 3, although AdaCLIP suppresses background interference effectively, its heatmaps show almost no response in most cases, leading to missed detections in insulator and metal plate categories. Tipsomaly, displayed in row 9, yields relatively cleaner backgrounds but sometimes generates diffused anomaly regions that do not tightly align with the actual anomaly boundaries. AdaptCLIP shown in row 10 maintains moderate localization ability but remains susceptible to background noise in complex scenes such as the fender hammer and triangular connection plates. The heatmaps of AF-CLIP in row 11 exhibit overly broad activation regions. While it identifies the general anomaly locations, the responses tend to cover extensive non anomaly areas, limiting the fine-grained precision required for industrial applications. In contrast, BAG-CLIP provides precise detection for challenging categories such as insulators, capsules, and pipes. This indicates that by propagating information through the graph structure, even regions with unobvious initial features can be enhanced by their neighboring salient anomaly nodes, thereby achieving robust detection of minute and morphologically complex anomalies.

4.3. Ablation Studies

4.3.1. Module Ablation Studies

To verify the effectiveness of the core components within the BAG-CLIP framework, a series of ablation experiments were conducted. These experiments primarily evaluated the contributions of two key modules: the SAG module and the BSA module. The effectiveness was assessed by removing these modules individually and testing on the four datasets, taking the average value. The results of the ablation experiments are presented in Table 3.

Table 3. Results of ablation experiment.

SAG	SAFR	Dataset	Image-Level			Pixel-Level		
			AUROC	F1-Max	AP	AUROC	F1-Max	AP
✓	✗	EPED	72.1	71.6	71.2	90.7	18.6	14.3
		InsPLAD	80.0	66.9	66.7	81.1	19.3	13.8
		MPDD	75.7	82.6	78.5	93.4	27.7	24.9
		BTAD	79.7	85.3	89.8	90.9	43.3	38.5
		Average	76.9	76.6	76.6	89.0	27.2	22.9
✗	✓	EPED	70.6	72.3	69.8	92.2	20.7	14.8
		InsPLAD	80.9	69.7	65.0	79.8	16.2	12.0
		MPDD	78.0	81.9	79.5	93.1	23.9	20.9
		BTAD	90.6	86.3	92.3	91.7	46.2	40.9
		Average	80.0	77.6	76.7	89.2	26.8	22.2
✓	✓	Average	84.3	82.9	82.3	96.1	35.3	30.9

Note: Red indicates the best result, "✓" indicates that the module is incorporated into the model, and "✗" indicates that it is not.

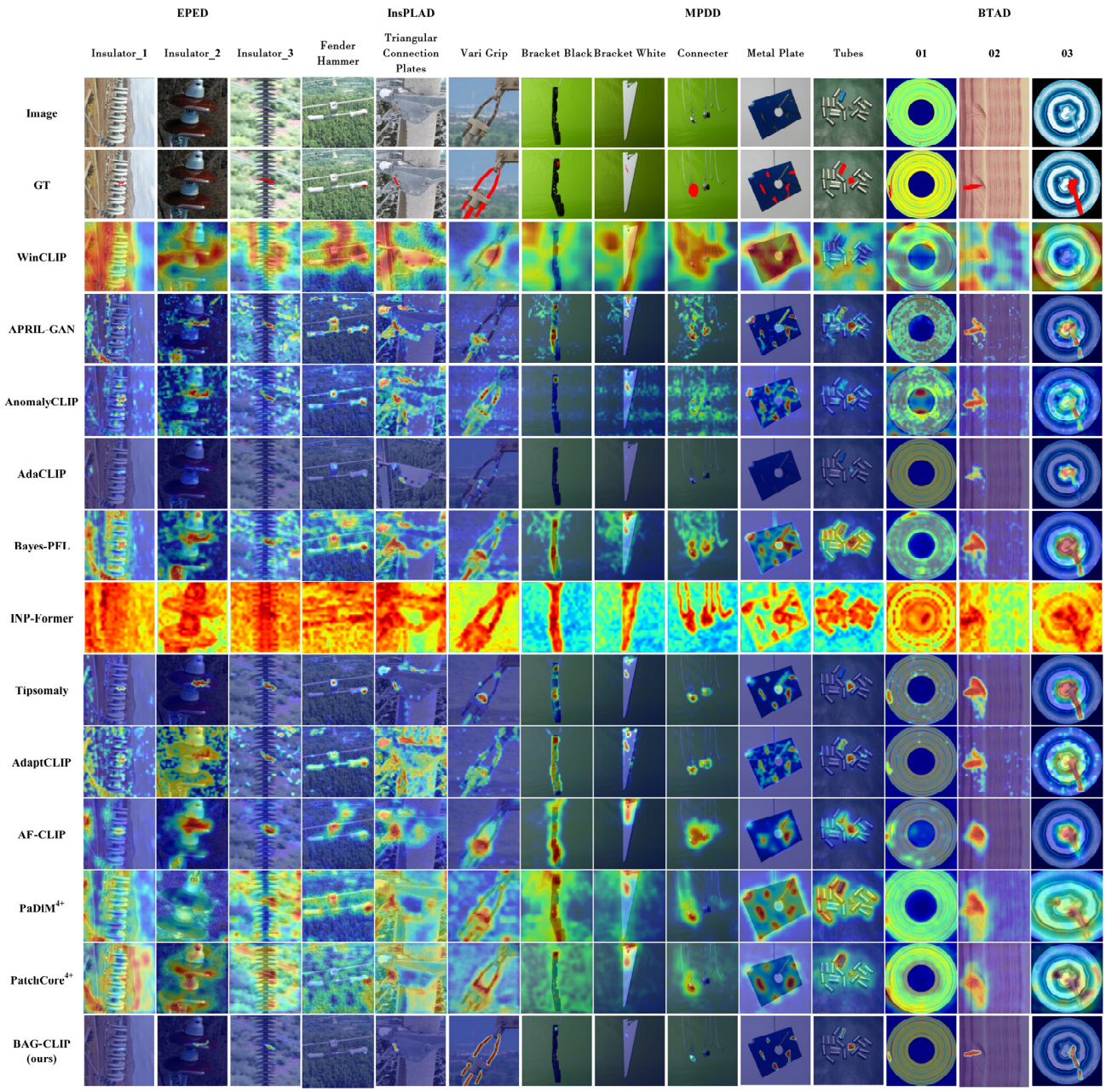

Figure 3. Visualization of anomaly maps of different anomaly detection methods.

When the SAFR module in the BSA module was removed, the model's metrics declined significantly. The image-level AUROC and F1 scores dropped from 84.3% and 82.9% to 76.9% and 76.6%, respectively, and the pixel-level AUROC dropped from 96.1% to 89.0%. By decoupling global semantics and spatial information, BAG-CLIP is able to preserve high-fidelity spatial features for pixel-level tasks, thereby achieving more precise anomaly localization.

When the feature enhancement SAG module was removed, the model's image-level AUROC decreased from 84.3% to 80.0%, and the pixel-level F1 score dropped drastically from 35.3% to 26.8%. This indicates that relying solely on the self-attention mechanism of ViT is insufficient to fully capture the local topological structure information crucial for anomaly detection. The SAG module explicitly models the contextual relationships between image patches by constructing a graph in the feature space and propagating information. This greatly enhances the localization and discrimination capability for anomalies with

complex morphologies, such as cracks and scratches, thereby significantly improving segmentation accuracy.

To visually verify the feature enhancement capabilities of the SAG module, this paper further employs heatmaps for visual analysis. As shown in Figure 4, without the SAG module, the heatmap exhibits a diffuse distribution, with noise responses present in the background areas, and fails to generate effective anomaly responses for anomalies such as bent insulators and rusted line clamps. In contrast, in the heatmap enhanced by the SAG module, cracks and bending in insulators are clearly detected, background noise is effectively suppressed, and the spatial continuity of the anomaly regions is significantly improved. These visualization results further validate that the SAG module constructs a graph structure in the feature space and aggregates neighborhood information, integrating the originally discontinuous anomaly responses into complete anomaly patterns, thereby providing more precise feature support for pixel-level anomaly localization.

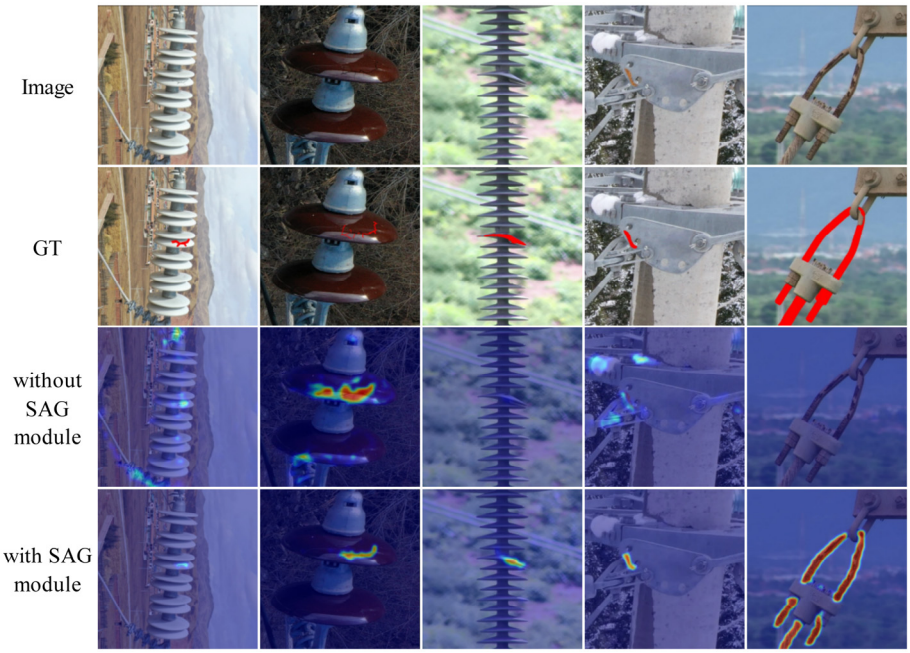

Figure 4. Visual comparison of anomaly heatmaps with and without the SAG module.

4.3.2. Key Parameter Sensitivity Analysis

The sensitivity analysis in this study focuses on two critical hyperparameters that dictate the structural modeling and information propagation within the SAG module: the number of neighbors K in the K-nearest neighbors algorithm, and the number of GCN layers P. To ensure a fair comparison with baseline methods, common training hyperparameters (e.g., learning rate, training epochs) follow typical settings from public literature and are excluded from the sensitivity analysis. The following sensitivity analysis experiments are conducted on both the EPED and MPDD datasets, and model performance is evaluated using the average values across all categories in the dataset.

The number of neighbors K determines the sparsity of the graph structure and the range of the local receptive field. As shown in Figure 5, when the value of K increases from 5 to 15, the various metrics of the model on both image-level and pixel-level tasks demonstrate a progressive upward trend across both datasets. The model achieves its optimal performance when K is set to 15. At this point, taking the EPED dataset as an example, the image-level AUROC and F1-Max reach 76.4% and 77.6%, respectively. This indicates that moderately increasing the number of neighbors helps the model aggregate richer local context information. However, when the value of K further increases to 20 or

30, the model's performance declines. This suggests that an excessively large neighborhood range introduces redundant noise irrelevant to the semantic meaning of the current node, leading to the dilution of effective features during the aggregation process, which in turn weakens the model's discriminative ability for minor anomalies.

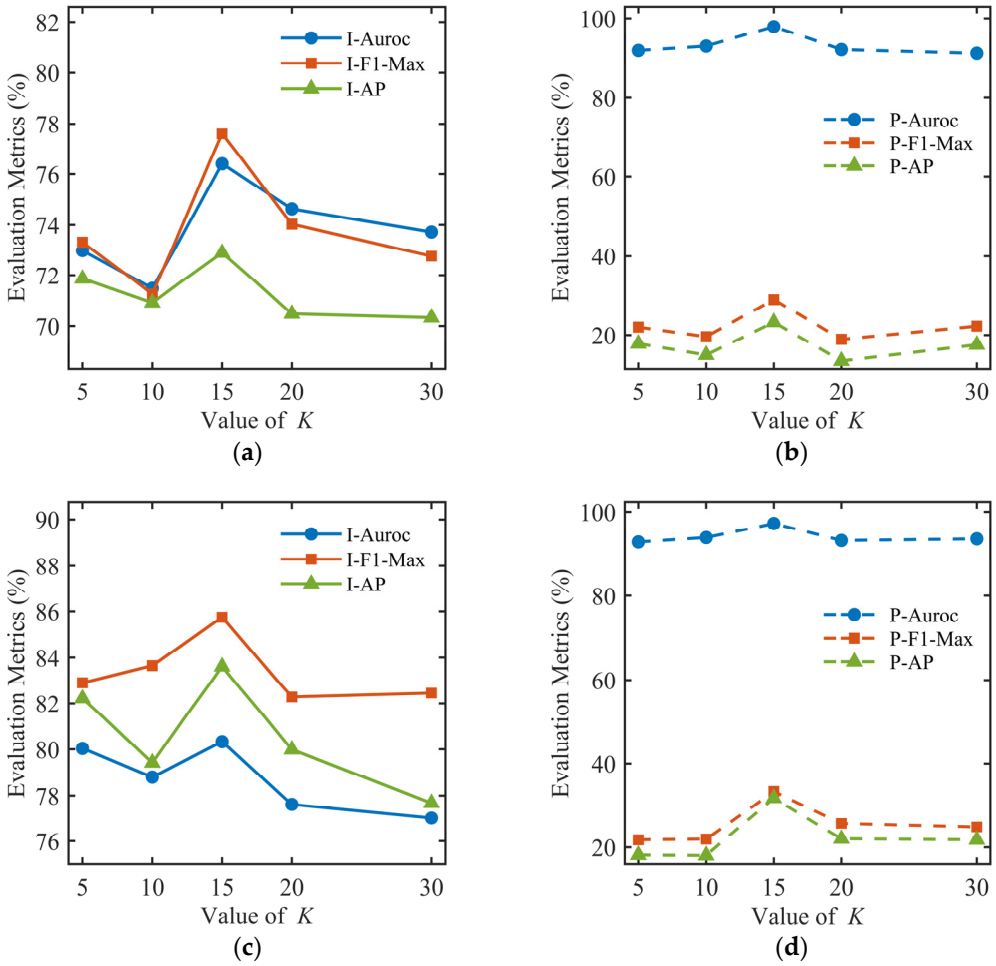

Figure 5. Impact of neighbor count K on detection performance. (**a**) The effect of the K value on image-level metrics on EPED dataset. (**b**) The effect of the K value on pixel-level metrics on EPED dataset. (**c**) The effect of the K value on image-level metrics on MPDD dataset. (**d**) The effect of the K value on pixel-level metrics on MPDD dataset.

The number of graph convolutional network layers P directly affects the depth and breadth of feature propagation within the graph structure. As shown in Figure 6, when the number of layers increases from 2 to 5, the detection accuracy improves with fluctuations, achieving optimal results at 5 layers, where the pixel-level AUROC on the EPED dataset reaches 97.9%. This indicates that a network structure of appropriate depth can facilitate high-order information interaction between nodes, which is beneficial for capturing complex topological structure features. Conversely, when the number of network layers continues to increase to 6 or 7, the performance metrics exhibit a significant decline. This occurs because excessive convolution operations cause the feature representations of different nodes to tend toward homogenization, blurring the feature boundaries between normal and anomalous regions and reducing the accuracy of anomaly detection.

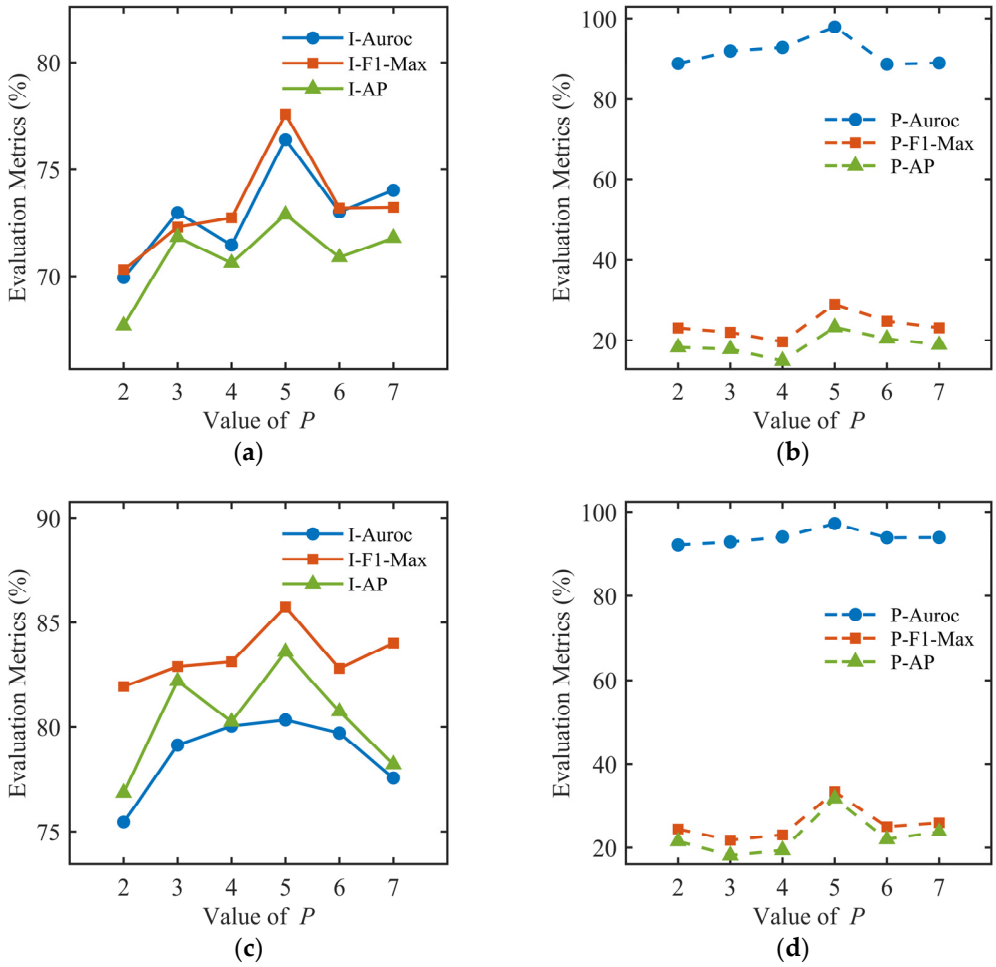

Figure 6. Impact of GCN layer P on detection performance. (**a**) The effect of the P value on image-level metrics on EPED dataset. (**b**) The effect of the P value on pixel-level metrics on EPED dataset. (**c**) The effect of the P value on image-level metrics on MPDD dataset. (**d**) The effect of the P value on pixel-level metrics on MPDD dataset.

4.4. Model Complexity Analysis

In real-world industrial anomaly detection, the precise identification of minute anomalies, such as micro-cracks or subtle surface scratches, heavily relies on high-resolution image inputs. To evaluate the engineering feasibility of the SAG module when processing high-resolution images, this paper conducts a systematic analysis of its computational complexity and theoretical overhead. Constructing a topological graph directly at the original pixel level would generate a massive number of nodes. This would lead to prohibitive computational and memory overheads.

Therefore, the SAG module deploys the dynamic graph construction process in the deep feature space. After feature extraction by the image encoder, the number of feature patches is effectively reduced to N. At this scale, the time complexity of calculating the Euclidean distance between nodes to generate the initial adjacency matrix is $O(N^2)$, and the spatial complexity is $O(N^2)$. This process relies solely on the similarity between features and does not require the introduction of additional learnable parameters. The complexity and computational overhead for each component of the SAG module are detailed in Table 4.

The experimental results demonstrate that a single inference pass of the SAG module takes only about 11 ms, proving that the proposed method effectively alleviates the curse of dimensionality in graph construction while achieving high-resolution feature extraction, thereby successfully balancing detection accuracy with real-time performance.

Table 4. Complexity and calculation overhead of SAG module.

Module Component	Parameters/M	Theoretical Computation/GFLOPs	Peak VRAM/MB	Single Inference Time/ms
Dynamic Graph Construction	None	3.84	67.4	2
GCN Feature Propagation	1.08	7.18	52.4	9
Total	1.08	11.02	67.4	11

To validate the engineering applicability of the proposed approach, we quantify model complexity using the total count of parameters (including both the frozen backbone and learnable components), and measure inference efficiency via single-image latency. The evaluation is performed on one NVIDIA A6000 48GB GPU. As presented in Table 5, all benchmarking methods employ the CLIP ViT-L/14@336px model with a unified input resolution of 518 × 518. Regarding model complexity, BAG-CLIP introduces only 13 M trainable parameters. In terms of inference efficiency, it attains a latency of 121.6 ms per image (8.22 FPS). Ultimately, BAG-CLIP strikes an excellent balance between the precise detection of complex morphological anomalies and efficient inference in industrial scenarios.

Table 5. Comprehensive comparison of inference efficiency.

Methods	Total Params/M	Inference Time/ms	FPS
AF-CLIP	$428.8 + 2.1 \times 10^0$	116.8	8.56
AnomalyCLIP	$428.1 + 6.2 \times 10^0$	123.8	8.08
AdaCLIP	$428.1 + 1.1 \times 10^1$	127.9	7.81
Bayes-PFL	$429.4 + 2.7 \times 10^1$	236.6	4.23
BAG-CLIP (Ours)	$429.0 + 1.3 \times 10^1$	121.6	8.22

4.5. Failure Cases

While the proposed BAG-CLIP framework performs robustly on diverse benchmark datasets, it still shows limitations under severe visual ambiguity caused by optical interference. As shown in Figure 7, intense specular reflections and irregular shadows on power equipment surfaces can visually mimic real anomaly features. This causes the model to misidentify illumination-induced pseudo-defects as anomalies, leading to false positives or missed subtle anomalies. Future work will focus on designing fine-grained text prompts with physical and lighting descriptions to improve cross-modal alignment and enhance the model's discriminability in confusing scenarios.

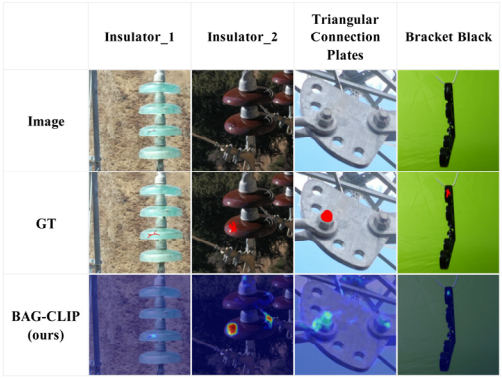

Figure 7. Failure cases of BAG-CLIP.

5. Conclusions

In this paper, we propose a novel Bifurcated Attention Graph-Enhanced CLIP framework, BAG-CLIP, to address the limited accuracy of existing vision-language model-based zero-shot anomaly detection methods when identifying complex morphological anomalies in transmission line inspection and industrial metal components. Firstly, a dual-path BSA module is introduced within the CLIP image encoder, which decouples feature learning into global semantic modeling and spatial detail preservation. This design effectively alleviates the inherent conflict between high-level semantic abstraction and fine-grained spatial localization, while providing high-quality feature representations for subsequent pixel-level anomaly segmentation. Secondly, a SAG module is developed to explicitly model the topological structures and contextual dependencies of complex industrial anomalies. Utilizing graph convolutional networks for information propagation, it enhances the model's representational capacity for complex morphological anomalies, thereby improving detection robustness. Finally, extensive experiments are conducted on multiple benchmark datasets covering electrical equipment, metal parts, and general industrial products. The experimental results demonstrate that BAG-CLIP significantly outperforms representative state-of-the-art zero-shot anomaly detection methods including WinCLIP, AnomalyCLIP, and AdaCLIP in both image-level classification and pixel-level segmentation. The proposed framework shows robust performance in detecting challenging complex morphological anomalies.

Author Contributions: Conceptualization, H.W., T.Z. and S.L.; methodology, H.W. and T.Z.; software, T.Z.; validation, H.W., T.Z. and S.L.; formal analysis, H.W., T.Z. and S.L.; investigation, H.W.; resources, T.Z.; data curation, T.Z.; writing—original draft preparation, H.W., T.Z. and S.L.; writing—review and editing, H.W., T.Z. and S.L.; visualization, H.W., T.Z. and S.L.; supervision, H.W.; project administration, H.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The EPED dataset will be made available on request. The other datasets used in this work are publicly available, MVTec AD: https://www.mvtec.com/company/research/datasets/mvtec-ad/downloads (accessed on 16 March 2026), VisA: https://github.com/amazon-science/spot-diff (accessed on 16 March 2026), MPDD: https://github.com/stepanje/MPDD (accessed on 16 March 2026), BTAD: https://github.com/dataset-ninja/btad (accessed on 16 March 2026), InsPLAD: https://github.com/andreluizbvs/InsPLAD (accessed on 16 March 2026), MIAD: https://miad-2022.github.io/ (accessed on 16 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VLMs	vision-language models
ZSAD	Zero-shot Anomaly Detection
ViT	Vision Transformer
GCN	Graph Convolutional Network
MLP	Multi-Layer Perceptron
AUROC	Area Under the Receiver Operating Characteristic Curve
F1-Max	maximum F1 score
AP	Average Precision
AUPRO	Area Under Per-Region Overlap
BSA	Bifurcated Self-Attention
SAG	Self-Attention Graph
SAFR	Scale-Adaptive Feature Recalibration
JFF	Joint Feature Fusion

References

1. Zhao, Y.; Liu, Q.; Su, H.; Zhang, J.; Ma, H.; Zou, W. Attention-based multiscale feature fusion for efficient surface defect detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5013310. [CrossRef]
2. Jha, S.B.; Babiceanu, R.F. Deep CNN-based visual defect detection: Survey of current literature. *Comput. Ind.* **2023**, *148*, 103911. [CrossRef]
3. Zhang, Z.; Chen, S.; Huang, J.; Ma, J. Zero-shot defect detection with anomaly attribute awareness via textual domain bridge. *IEEE Sens. J.* **2025**, *25*, 11759–11771. [CrossRef]
4. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; pp. 8748–8763.
5. Ma, W.; Zhang, X.; Yao, Q.; Tang, F.; Wu, C.; Li, Y.; Yan, R.; Jiang, Z.; Zhou, S.K. Aa-CLIP: Enhancing zero-shot anomaly detection via anomaly-aware CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 4744–4754. [CrossRef]
6. Liu, Y.; Li, Q.; Wang, Z.; Kato, J.; Zhang, J.; Wang, W. LECLIP: Boosting zero-shot anomaly detection with local enhanced CLIP. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5034111. [CrossRef]
7. Wu, H.; Jia, D.; Zhang, T.; Bai, X.; Sun, L.; Pu, M. Multimodal zero-shot anomaly detection using dual-experts for electrical power equipment inspection images. *J. Image Graph.* **2025**, *30*, 672–682. [CrossRef]
8. Vieira e Silva, A.L.B.; de Castro Felix, H.; Simões, F.P.M.; Teichrieb, V.; dos Santos, M.; Santiago, H.; Sgotti, V.; Lott Neto, H. InsPLAD: A dataset and benchmark for power line asset inspection in UAV images. *Int. J. Remote Sens.* **2023**, *44*, 7294–7320. [CrossRef]
9. Jezek, S.; Jonak, M.; Burget, R.; Dvorak, P.; Skotak, M. Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions. In Proceedings of the 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, 25–27 October 2021; pp. 66–71. [CrossRef]
10. Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; Foresti, G.L. VT-ADL: A vision transformer network for image anomaly detection and localization. In Proceedings of the IEEE 30th International Symposium on Industrial Electronics (ISIE), Kyoto, Japan, 20–23 June 2021; pp. 1–6. [CrossRef]
11. Bao, T.; Chen, J.; Li, W.; Wang, X.; Fei, J.; Wu, L.; Zhao, R.; Zheng, Y. MIAD: A maintenance inspection dataset for unsuper-vised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Paris, France, 2–6 October 2023; pp. 993–1002. Available online: https://ieeexplore.ieee.org/document/10350876 (accessed on 16 March 2026).
12. Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; Dabeer, O. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19606–19616. [CrossRef]
13. Zhou, Q.; Pang, G.; Tian, Y.; He, S.; Chen, J. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024; pp. 49705–49737.
14. Qu, Z.; Tao, X.; Gong, X.; Qu, S.; Chen, Q.; Zhang, Z. Bayesian prompt flow learning for zero-shot anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 30398–30408. [CrossRef]
15. Luo, W.; Cao, Y.; Yao, H.; Zhang, X.; Lou, J.; Cheng, Y. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 9974–9983. [CrossRef]
16. Duan, M.; Mao, L.; Liu, R.; Liu, W.; Liu, Z. Unified model based on reinforced feature reconstruction for metro track anomaly detection. *IEEE Sens. J.* **2024**, *24*, 5025–5038. [CrossRef]
17. Xiang, P.; Ali, S.; Jung, S.K.; Zhou, H. Hyperspectral anomaly detection with guided autoencoder. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5538818. [CrossRef]
18. Xia, X.; Pan, X.; Li, N.; He, X.; Ma, L.; Zhang, X.; Ding, N. GAN-based anomaly detection: A review. *Neurocomputing* **2022**, *493*, 497–535. [CrossRef]
19. Pang, G.; Shen, C.; Cao, L.; van den Hengel, A. Deep learning for anomaly detection: A review. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
20. Zhou, Y.; Liang, X.; Zhang, W.; Zhang, L.; Song, X. VAE-based deep SVDD for anomaly detection. *Neurocomputing* **2021**, *453*, 131–140. [CrossRef]
21. Zhang, Z.; Deng, X. Anomaly detection using improved deep SVDD model with data structure preservation. *Pattern Recognit. Lett.* **2021**, *148*, 1–6. [CrossRef]
22. Li, Z.; Yan, H.; Tsung, F.; Zhang, K. Profile decomposition based hybrid transfer learning for cold-start data anomaly detection. *ACM Trans. Knowl. Discov. Data* **2022**, *16*, 121. [CrossRef]

23. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14318–14328. [CrossRef]
24. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges, Virtual Event, Milan, Italy, 10–15 January 2021; pp. 475–489. [CrossRef]
25. Chen, X.; Han, Y.; Zhang, J. 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* **2023**, arXiv:2305.17382. [CrossRef]
26. Cao, Y.; Zhang, J.; Frittoli, L.; Cheng, Y.; Shen, W.; Boracchi, G. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In Proceedings of the 18th European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 55–72. [CrossRef]
27. Zhu, H.; Zhao, C.; Yuan, Y.; Liu, M. A zero-shot anomaly detection network with patch-augmented prompts and test-time adaptation. *Eng. Appl. Artif. Intell.* **2026**, *165*, 113525. [CrossRef]
28. Chen, P.; Huang, F.; Huang, C. DyC-CLIP: Dynamic context-aware multi-modal prompt learning for zero-shot anomaly detection. *Pattern Recognit.* **2026**, *176*, 113215. [CrossRef]
29. Kim, D.; Park, C.; Cho, S.; Lim, H.; Kang, M.; Lee, J.; Lee, S. Generalizing CLIP prompts for zero-shot anomaly detection. *Pattern Recognit.* **2026**, *178*, 113406. [CrossRef]
30. Chen, X.; Zhang, J.; Tian, G.; He, H.; Zhang, W.; Wang, Y.; Wang, C.; Liu, Y. CLIP-AD: A language-guided staged dual-path model for zero-shot anomaly detection. In Proceedings of the Human Activity Recognition and Anomaly Detection (IJCAI 2024), Jeju, Republic of Korea, 3–9 August 2024; pp. 17–33. [CrossRef]
31. Gao, B.B.; Zhou, Y.; Yan, J.; Cai, Y.; Zhang, W.; Wang, M.; Liu, J.; Liu, Y.; Wang, L.; Wang, C. AdaptCLIP: Adapting CLIP for Universal Visual Anomaly Detection. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Singapore, 2026. Available online: https://api.semanticscholar.org/CorpusID:278636461 (accessed on 16 March 2026).
32. Fang, Q.; Lv, W.; Su, Q. AF-CLIP: Zero-Shot Anomaly Detection via Anomaly-Focused CLIP Adaptation. In Proceedings of the ACM International Conference on Multimedia (ACM MM), 2025. Available online: https://api.semanticscholar.org/CorpusID:280322959 (accessed on 16 March 2026).
33. Salehi, M.R.; Sadjadi, N.; Baselizadeh, S.; Rabiee, H.R. TIPS Over Tricks: Simple Prompts for Effective Zero-Shot Anomaly Detection. *arXiv* **2026**, arXiv:2602.03594. [CrossRef]
34. Gao, B.-B.; Wang, C.J. One Language-Free Foundation Model Is Enough for Universal Vision Anomaly Detection. *arXiv* **2026**, arXiv:2601.05552. Available online: https://api.semanticscholar.org/CorpusID:284597414 (accessed on 16 March 2026). [CrossRef]
35. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. Available online: https://link.springer.com/chapter/10.1007/978-3-030-01234-2_1 (accessed on 16 March 2026).
36. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
37. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516. Available online: https://arxiv.org/abs/1911.09516 (accessed on 16 March 2026). [CrossRef]
38. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 4700–4708. [CrossRef]
39. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
40. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]
41. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. MVTec AD—A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9584–9592. [CrossRef]
42. Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; Dabeer, O. SPot-the-Difference Self-supervised Pre-training for Anomaly Detection and Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 392–408. [CrossRef]