Self-Supervised CLIP-Guided for Few-Shot Industrial Anomaly Detection

Yingwen Chen, Ying Xu, Tianlei Wang, Yikui Zhai, *Senior Member, IEEE,* Kanghong Tan, Jianhong Zhou, Pasquale Coscia, *Senior Member, IEEE,* Angelo Genovese, *Senior Member, IEEE,* C. L. Philip Chen, *Life Fellow, IEEE*

***Abstract*—Few-shot industrial anomaly detection aims to identify unseen defects using only a limited number of normal samples. However, most existing approaches still rely heavily on auxiliary industrial datasets for training. In this paper, we propose a novel self-supervised CLIP-guided for few-shot industrial anomaly detection, which eliminates the need for auxiliary industrial data. Specifically, we first introduce a pseudo-anomaly generation strategy that synthesizes both structural and textural anomalies. Then, leveraging the cross-modal semantic understanding capability of CLIP, we contrast the multi-scale visual features with learnable textual prompts to achieve anomaly localization grounded in language semantics. Inspired by the human cognitive process of identifying anomalies through reference comparison, we introduce a support set composed of a few normal samples and perform semantic-level feature alignment with the query set via CLIP visual encoder, thereby enhancing anomaly discrimination. Furthermore, we also introduce Adapter to alleviate the semantic offset problem between text and image modalities in industrial scenarios of CLIP, and enhance the model's robustness to the spatial structure differences between query set and support set. Extensive experiments conducted on the MVTec AD, the VisA, the BTAD and the MPDD datasets demonstrate that our method achieves competitive results under the few-shot setting. Moreover, its effectiveness and deployability are validated through real-world application in battery spot-welding defect inspection. The code is available at https://github.com/YiKuiZhai/SCF-AD.**

***Index Terms*—CLIP, cross-modal, representation alignment, industrial anomaly detection, few-shot.**

I. Introduction

This study was funded by Guangdong Higher Education Innovation and Strengthening School Project (No. 2020ZDZX3031, No. 2022ZDZX1032, No. 2023ZDZX1029), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology Research Project (No. 2220002000246, No. 2023760300070008390). This work was supported by the High Performance Computing Center of Wuyi University. (Yingwen Chen, Ying Xu, Tianlei Wang, Yikui Zhai, Kanghong Tan, Jianhong Zhou contributed equally to this work.) (Corresponding author: Yikui Zhai.)

Yingwen Chen, Ying Xu, Tianlei Wang, Yikui Zhai and Kanghong Tan are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, 529020, China (e-mail: yingwenchen2024@163.com; xuying117@163.com; tianlei.wang@aliyun.com; yikuizhai@163.com; tankanghong234@163.com).

Jianhong Zhou are with the School of Electronics and Information Engineering, South China University of Technology, Guangzhou, Guangdong 510641, China (e-mail: jackychou_lab@126.com).

Pasquale Coscia, Angelo Genovese are with the Department of Computer Science, Università Degli Studi di Milano, 20133, Milan, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it).

C. L. Philip Chen is with the Faculty of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510006, China (e-mail: philip.chen@ieee.org).

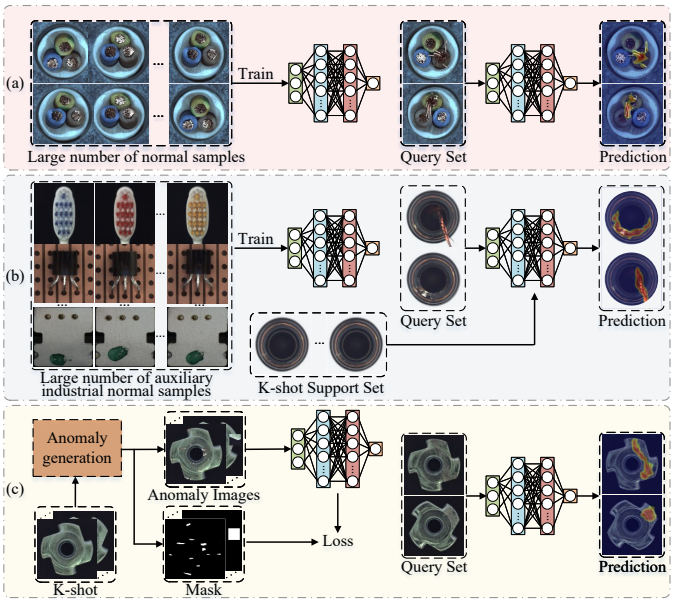

Fig. 1. Comparing the training and inference pipelines of SCF-AD and mainstream anomaly detection paradigms. (a) Vanilla anomaly detection. (b) Few-shot anomaly detection with auxiliary industrial dataset. (c) Self-supervised few-shot anomaly detection without auxiliary industrial dataset.

IN industrial measurement, accurate anomaly detection is crucial for ensuring product quality and system reliability by identifying rare and unforeseen abnormal states from predominantly normal data streams [1]–[3]. Traditional defect detection methods, such as SIFT-based handcrafted features [4], can be integrated with Automated Optical Inspection (AOI) systems [5] but suffer from limited and cumbersome feature extraction, making them ineffective in dynamic environments and against unseen anomalies, thus lacking generalization.

In contrast, deep learning–based approaches can be integrated with vision measurement devices to perform tasks such as defect classification [6], segmentation [7], or object detection [8], but these methods usually rely on a large number of labeled defective samples for supervised training. This dependence limits their practicality, as abnormal instances are often extremely scarce and expensive to annotate in real-world industrial settings [9]. Industrial anomaly detection, however, enables the detection and localization of unknown anomalies using only normal samples, offering superior generalization and deployment flexibility. These advantages have made it

an increasingly important research focus in industrial visual inspection [10].

Prior studies have broadly categorized industrial anomaly detection approaches into three main types. i) Reconstruction-based methods [11], [12] employ autoencoders or generative adversarial networks to regenerate normal images and identify anomalies by analyzing reconstruction discrepancies. ii) Feature comparison techniques [13]–[15] extract high-level semantic representations and assess their similarity to a reference set in order to determine deviations from normal patterns. iii) Anomaly generation-based methods [16]–[19] generate pseudo-anomalous samples, guiding the model to autonomously learn discriminative representations for distinguishing between normal and abnormal patterns. As illustrated in Fig. 1(a), the above methods typically require a substantial number of normal samples during training. However, in practical industrial applications, acquiring sufficient high-quality data for each new product is time-consuming and significantly hinders deployment efficiency. Furthermore, such datasets often exhibit high redundancy, meaning that collecting excessive amounts of normal data yields diminishing returns in performance while imposing additional costs in terms of storage and preprocessing. To address the aforementioned challenges, few-shot industrial anomaly detection (FSAD) has attracted increasing research interest in recent years. These approaches typically rely on only a few reference images (e.g., 1-8 normal samples) during inference to detect and localize anomalies. Representative approaches such as RegAD [20] and ADFormer [21] incorporate normal images from auxiliary categories during training to capture general semantic knowledge. During inference, only a small number of target-category images are used as references, mimicking human behavior of identifying anomalies through comparison with normal samples to achieve precise defect recognition. However, these methods rely solely on visual inputs and overlook the potential of textual semantics.

With the strong transferability demonstrated by vision-language pre-trained models such as Contrastive language-image pretraining (CLIP) [22] in generic image understanding tasks, an increasing number of researchers have begun investigating their applications in downstream tasks. Significant advances have been achieved in areas such as defect classification [23], anomaly detection [24]–[26], and text recognition [27]. CLIP aligns visual and linguistic features within a shared semantic space by performing cross-modal contrastive learning on approximately 400 million image-text pairs. This enables it to acquire strong semantic understanding and category generalization capabilities, offering a new solution to few-shot learning challenges [23], [28], [29]. Recently, researchers have begun integrating CLIP into FSAD frameworks, with methods like FADE [30], APRIL-GAN [31], and InCTRL [32] all relying on CLIP's image-text contrastive mechanism, leveraging language guidance and visual references to construct a unified anomaly-aware representation space. Despite their strong performance in FSAD settings, these methods still depend on external auxiliary-category data during training. As shown in Fig. 1(b), these approaches essentially follow the paradigm of "auxiliary industrial data + few-shot reference matching", thereby exhibiting a degree of data dependence. It is worth noting that AnomalyCLIP [33], although designed for zero-shot anomaly detection, still requires auxiliary-domain data containing defective samples for fine-tuning. In real-world applications, constructing high-quality auxiliary industrial datasets can be prohibitively expensive. Although CLIP offers general-purpose visual and textual representations for FSAD, it still encounters semantic misalignment issues in industrial contexts, as it is pre-trained on open-domain image–text pairs and lacks awareness of fine-grained industrial anomaly concepts. This leads to a modality shift between visual and textual representations. In addition, spatial structural differences between the query and support sets of the same product—such as variations in viewpoint, positioning, or deformation—often arise during inference and further impair semantic alignment. In addition, recent studies have demonstrated that self-supervised, multimodal, and few-shot [34]–[36] learning strategies can effectively enhance detection and representation capabilities under limited data, further validating the potential of cross-modal perception frameworks in industrial scenarios [37].

To address the aforementioned challenges, we propose a novel self-supervised CLIP-guided few-shot industrial anomaly detection (SCF-AD) method that eliminates the need for auxiliary industrial data, as illustrated in Fig. 1(c). This approach relies solely on a small number of normal samples from the target category for model training and employs pseudo-anomaly generation to facilitate the learning of anomaly perception capabilities. Specifically, the framework incorporates a language-aware visual alignment and a human perceptual contrast matching branches to capture semantic correspondence from textual and visual references. Both branches operate on intermediate representations of the query image extracted from multiple layers of the CLIP visual encoder (CLIP-V), computing semantic similarity with text embeddings and support set image features, enabling the model to distinguish anomalies across different semantic levels. Furthermore, we introduce Adapters into both the vision-language alignment and the query-support comparison branches to alleviate the modality shift of CLIP under industrial semantics and to enhance the model's adaptability to structural variations. The main contributions of this work are as follows:

1) A novel self-supervised CLIP-guided few-shot industrial anomaly detection framework, named SCF-AD, is proposed without relying on auxiliary industrial datasets. To the best of our knowledge, this is the first paradigm in this context to perform self-supervised training using only a few normal samples.
2) A language-aware visual alignment branch and a human perceptual contrast matching branch are incorporated to capture semantic correspondence from textual and visual references. An Adapter module is further designed to alleviate modality discrepancies between CLIP Vision and CLIP Text and enhance robustness to spatial variations between query and support images.
3) The proposed method was extensively evaluated on the MVTec AD, the VisA, the BTAD, and the MPDD

datasets. The results demonstrate that it achieves highly competitive performance compared to existing state-of-the-art methods. Furthermore, its effectiveness was validated in a real-world battery spot-welding defect detection scenario.

The rest of this paper is structured as follows. Section II surveys the literature on industrial anomaly detection, focusing on both traditional techniques and the latest progress in few-shot learning methods. Section III provides a comprehensive explanation of the proposed SCF-AD framework. Section IV presents the experimental section, detailing the datasets, evaluation metrics, experimental setup, ablation studies, and comparative experiments. Section V concludes the paper and discusses potential directions for future research.

II. RELATED WORKS

A. Vanilla Industry Anomaly Detection

Reconstruction-based. These methods identify and localize anomalous regions by learning to reconstruct only normal images and measuring differences between the original and reconstructed images. Li et al. [11] developed a method based on an Adversarial Autoencoder, in which the first-stage reconstructed image is re-fed into the model for a second reconstruction to produce clearer images. Zhang et al. [12] proposed a method that restores multi-level semantic features and compares residuals to locate multi-scale detailed defects. Although reconstruction-based methods perform well in anomaly detection, they struggle to accurately recover defects with low contrast or those difficult to distinguish from the background.

Embedding-based. These methods extract features of normal samples using pretrained networks and detect anomalies by measuring the feature distance to test samples. These methods mainly fall into several paradigms: One paradigm is memory bank-based methods, such as PatchCore proposed by Roth et al. [13], which build a memory bank of normal features, compress it using coreset sampling, and locate anomalies via nearest neighbor search. Another category is knowledge distillation-based methods, where a student network is trained to imitate the teacher network's normal feature outputs, detecting anomalies through feature discrepancies. However, the conventional symmetric teacher-student network architecture limits the ability to express anomalies. To address this, Zhu et al. [14] proposed an asymmetric structure to enhance the expression of anomaly feature differences between teacher and student networks. Furthermore, flow-based methods map complex normal feature distributions to simple prior distributions through a series of invertible transformations, identifying anomalies via low likelihood values. However, conventional flow models struggle to handle anomalies with large variations in scale. To address this, Zhou et al. [15] proposed MSFlow, a multi-scale framework employing asymmetric parallel flow structures for multi-scale feature processing, and introduced fusion flows for information exchange, thereby better generalizing to anomalies of various sizes.

Anomaly generation-based. These methods aim to compensate for the absence of real anomaly samples by designing proxy tasks. The mainstream idea is to synthesize pseudo-anomalies on normal images and train models to distinguish normal from anomalous regions. Li et al. [16] proposed CutPaste, which creates structural anomalies by cutting and pasting patches within the same image, thereby disrupting its original structure. Zavrtanik et al. [17] proposed DRAEM, which uses Perlin noise to generate irregular masks for anomalous regions and randomly samples patches from an external texture dataset as anomaly textures, overlaying both onto normal training images to create pseudo-anomalies. Cao et al. [18], in their CDO method, add simple random noise patches onto normal images to aid training, aiming to better separate normal and anomalous feature distributions in feature space and alleviate model overgeneralization. Chen et al. [19] proposed U²D²PCB, which directly extracts defect regions from limited real defect samples and synthesizes them onto defect-free images to create more realistic training data. Zhai et al. [38] proposed an innovative framework for container marking anomaly detection, which employs a Cellular Anomaly Generation Strategy to synthesize pseudo-anomalous samples resembling real images and incorporates a Siamese network to detect anomalies based on inter-image differences. In this work, we integrate DRAEM and variants of CutPaste to respectively simulate texture and structural anomalies in real-world scenarios, providing new insights for self-supervised training under few-shot conditions.

B. Few-shot Industry Anomaly Detection

The aforementioned conventional anomaly detection methods generally rely on training with a large number of normal samples, as illustrated in Fig. 1(a). This results in high data collection costs and low deployment efficiency in modern industries with rapid product model iterations. To address this challenge, researchers have explored FSAD. For instance, PatchCore [13], as previously mentioned, demonstrated robust performance even under few-shot settings. Huang et al. [20] proposed RegAD, which uses a cross-category image registration network to align features between query images and normal support set images, detecting anomalies based on feature discrepancies. Zhu et al. [21] developed ADFormer, which features a dual CNN-Transformer backbone to capture both local and global features, and employs self-supervised bipartite matching to reconstruct query images from normal support images, using reconstruction errors to localize anomalies.

Recently, the emergence of large multimodal models [1], [35], [39], [40] with robust generalization capabilities has opened up new avenues for few-shot anomaly detection (FSAD) research. Some approaches leverage CLIP's strong vision-language alignment ability, using textual prompts to guide the model in understanding the semantics of "normal" and "anomalous". Jeong et al. [41] proposed WinCLIP, which enhances CLIP's ability to localize local anomalies via a windowing strategy and compositional prompt engineering. Li et al. [30] presented FADE, incorporating a language-sensitive Grounding Everything Module to further improve image-text alignment, while automatically generating prompts with large language method to better capture the concepts of anomalies and normality. Chen et al. [31] proposed APRIL-GAN, which

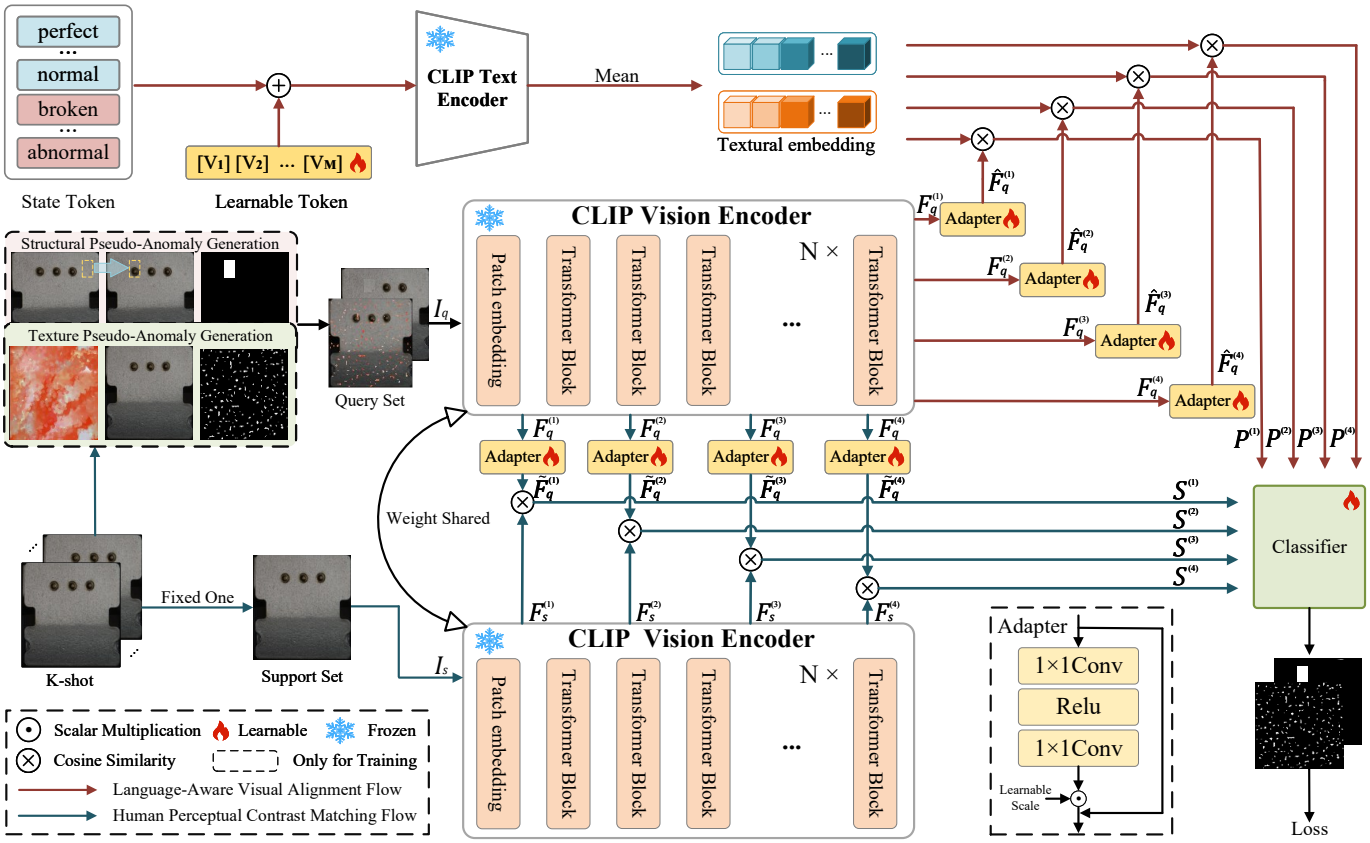

Fig. 2. The proposed self-supervised few-shot anomaly detection framework. It comprises pseudo- anomaly generation, learnable text prompts, Adapters, a Decoder, and frozen CLIP-V and CLIP-T. The pseudo-anomaly generation module is removed during inference. The framework operates solely on a few-shot set of target-class normal images for training, without requiring any auxiliary industrial data.

adds learnable linear layers after multiple stages of the image encoder to effectively map multi-scale image features into a shared vision-language space, addressing the issue that CLIP's intermediate features cannot be directly compared with text features. Zhu and Pang [32] proposed InCTRL, designed to learn residuals between query images and normal sample prompts, which not only compares their feature differences visually but also integrates semantic prior knowledge guided by text prompts for few-shot anomaly detection. Gu et al. [42] took this further with AnomalyGPT, leveraging the reasoning capability of large vision-language models and prompt fine-tuning, enabling the model not only to localize anomalies but also to diagnose and describe anomalies conversationally.

However, most of the aforementioned methods [20], [21], [31], [32], [42] follow the "auxiliary data + few-shot reference matching" paradigm illustrated in Fig. 1(b), relying on large-scale external datasets for training. For example, RegAD and ADFormer employ a "leave-one-out" training protocol on the MVTec AD [43], where data from the other 14 categories is used as auxiliary training when testing a specific category. Meanwhile, CLIP-based methods like APRIL-GAN and InCTRL require training on large auxiliary industrial datasets such as VisA [44] when tested on the MVTec AD dataset. Recently, methods such as AnomalyDINO [45] and PromptAD [46] have also emerged that avoid these auxiliary industrial datasets. These approaches, however, primarily employ the k-shot normal samples to construct a reference memory bank for feature matching. In particular, PromptAD constructs Manual Abnormal Prompts by relying on object-customized manual anomaly suffixes, which implies a dependency on prior expert knowledge regarding specific defect types. In contrast, as depicted in Fig. 1(c), the method proposed in this work explores a novel paradigm that requires no external auxiliary industrial data and trains the model solely via self-supervised learning on k-shot normal samples from the target category for anomaly detection through pseudo-anomaly generation.

III. PROPOSED METHOD

A. Framework Overview

Current FSAD methods generally rely on additional training data to learn generalized semantic features, but such auxiliary industrial data is often difficult to obtain in industrial settings, and the semantic shift across different categories limits the model's transferability. Meanwhile, although the CLIP possesses strong vision-language alignment capabilities, it still suffers from modality shift and spatial structural discrepancies in industrial domains, which significantly impact anomaly perception performance. To address this, we propose SCF-AD, a self-supervised CLIP-guided few-shot framework for industrial anomaly detection, which enables accurate detection and localization using only a limited number of normal samples, without relying on external auxiliary industrial data.

SCF-AD comprises four key components: pseudo-anomaly generation, semantic fusion and decoding, language-aware visual alignment, and human perceptual contrast matching. The latter two are designed as parallel semantic pathways, modeling the similarity between the query image and either language priors or reference images, as illustrated in Fig. 2. Meanwhile, Adapters are incorporated into both branches to enhance semantic alignment and structural robustness.

B. Pseudo-anomaly generation

Although CLIP possesses powerful zero-shot classification capabilities, it lacks the pixel-level granularity required for defect localization. Therefore, we introduce a pseudo-anomaly generation strategy to guide CLIP's attention from global semantic alignment to local structural perception. Specifically, in industrial visual inspection scenarios, anomalies are generally categorized into two types: i) one characterized by local texture disturbances, such as scratches, stains, or spots; ii) the other by structural deformations, missing regions, or spatial dislocations. Based on this observation, we introduce three types of data into the self-supervised training: normal images (Normal), which serve as reference samples for learning normal patterns; pseudo-texture anomalies (Texture), which are synthesized by simulating local texture disturbances; and pseudo-structural anomalies (Structure), which simulate macroscopic structural defects such as occlusion or misalignment.

Specifically, the generation of pseudo-texture anomalies follows the approach proposed in [17], where a training image I is first sampled from the k-shot normal images of the target class, then a texture background is generated using Perlin noise, from which a binary mask M is derived by applying a threshold. Meanwhile, an image I_s is randomly selected from the Describable Textures Dataset (DTD) [47] to serve as the anomaly source image. Finally, the generation of the pseudo-anomalous image I_q can be defined as follows:

$$I_q = \bar{M} \odot I + \alpha \left(M \odot I_s\right) + (1 - \alpha) \left(M \odot I\right) \tag{1}$$

Where $\bar{M}$ denotes the complement of the mask M, $\odot$ indicates element-wise multiplication, and a is a random scalar serving as an opacity factor to control the blending ratio between I and I_s in the anomalous region. However, for displacement-sensitive categories with structural defects (e.g., the Transistor in the MVTec dataset), simple texture noise is insufficient to simulate realistic anomalies. To enhance the model's robustness to spatial perturbations, an enhanced version of CutPaste [16] disrupts spatial consistency by randomly copying and pasting patches within an image, and we adopt this improved version [48], [49], further utilizing Poisson image editing [50] for seamless image blending, thereby constructing pseudo-anomalous images with local structural perturbations, which helps the model better capture geometric variations of non-textural anomalies. During training, we retain the original image (as a normal sample) with a probability of 30%, apply the texture anomaly generation with a 50% probability, and utilize the structure anomaly generation with a 20% probability. During the testing phase, the anomaly generation is removed, and the query image is directly input into the CLIP-V.

The pseudo-anomaly generation strategy does not aim to perfectly replicate the complex distribution of real-world defects. Instead, the objective is to achieve "Outlier Exposure". By presenting the model with a diverse set of synthetically generated structural and textural "outliers", the model is compelled to learn a highly compact and stable feature representation that exclusively defines normality. The core capability that generalizes to detecting real, unseen anomalies stems from this process, as it enables the model to identify significant deviation from this learned normal representation.

C. Language-Aware Visual Alignment

In early CLIP-based industrial anomaly detection methods, such as WinCLIP [41], fixed templates (e.g., "a photo of a [class]") were typically used to construct textual prompts. Such static templates are inadequate for capturing the fine-grained semantic differences between "normal" and "anomalous" in industrial settings. As a result, their ability to align with visual features of the query image is limited. Inspired by the success of CoOp in few-shot classification, the prompt-learning strategy with learnable context vectors [51] is incorporated to enhance CLIP's adaptability under limited data conditions. Although CLIP is pretrained on natural language, the first M context vectors in our method do not correspond to specific words, but are continuous learnable parameters optimized through training, which can form more discriminative implicit semantic prompts in the feature space, thereby enhancing the alignment between "normal" and "anomalous" states, finally, the textual prompt is formulated as follows:

$$g_n = [V_1, V_2, \ldots, V_M, \text{Normal\_State}] \tag{2}$$

$$g_a = [W_1, W_2, \ldots, W_M, \text{Abnormal\_State}] \tag{3}$$

Where, V_i and W_i denote the i^{th} learnable token for the "normal" and "abnormal" states, respectively, and M represents the number of learnable tokens. Normal_State and Abnormal_State represent the embeddings of non-learnable state category words, such as using "flawless" to denote the normal state and "broken" to indicate the abnormal state. The text embeddings of these state words are averaged to obtain the final semantic vectors for each state, collectively denoted as $E_{NA} \in \mathbb{R}^{C \times 2}$, which includes the semantic embeddings for both the "normal" and "abnormal" states.

The language-aware visual alignment is designed to guide visual features using text prompts, achieving semantic perception and anomaly detection in image regions via cross-modal semantic alignment. To obtain the textual priors, the pre-computed semantic vectors E_{NA} are passed through a frozen CLIP text encoder (CLIP-T):

$$F_t = \text{CLIP-T}\left(E_{NA}\right) \in \mathbb{R}^{C \times 2} \tag{4}$$

where C is the feature dimension of the CLIP-T output, and 2 corresponds to the two semantic categories: "normal" and "abnormal". Given a query image I_q, its features are extracted from the i^{th} layer of the CLIP-V as:

$$F_q^{(l)} = \text{CLIP-V}^{(l)}(I_q) \in \mathbb{R}^{B \times C \times H \times W} \tag{5}$$

Where I_q denotes the input query image, B represents the batch size, C is the number of channels, and H and W are the height and width of the feature map, respectively. Since the CLIP-V is pretrained on open-domain images, the extracted image features often exhibit semantic shift in industrial defect scenarios, making it difficult t o a ccurately a lign w ith fine-grained textual semantics. To address this issue, an Adapter$_t$ is introduced before computing similarity between image and text features, which semantically adapts the multi-scale intermediate features of the query image. As illustrated in Fig. 2, the Adapter adopts a lightweight bottleneck architecture composed of two sequential 1×1 convolutional layers with a ReLU activation and a residual connection. The residual design preserves the strong pre-trained CLIP features, while the learnable components focus on refining s emantic alignment, thereby effectively mitigating the cross-modal shift with minimal computational overhead. The overall process is as follows:

$$\hat{F}_q^{(l)} = \text{Adapter}_t^{(l)}\left(F_q^{(l)}\right) \in \mathbb{R}^{B\times C\times H\times W} \tag{6}$$

$$P^{(l)} = \text{Sim}\left(\hat{F}_q^{(l)}, F_t\right) = \sigma\left(\hat{F}_q^{(l)} \cdot F_t\right) \in \mathbb{R}^{B\times 2\times H\times W} \tag{7}$$

Where $\hat{F}_q^{(l)}$ denotes the semantically enhanced query image features, $P^{(l)}$ represents the language-aware similarity map indicating the degree to which each spatial position corresponds to the "normal" or "anomalous" semantic states, and $\sigma(\cdot)$ denotes the softmax operation applied along the channel dimension to normalize the similarity responses.

D. Human Perceptual Contrast Matching

In human perception, anomaly detection commonly relies on empirical understanding of the "normal state". Especially in industrial visual tasks, operators usually identify potential anomalies by comparing the current query image against a standard normal sample. This "reference-contrast" paradigm is highly generalizable and practical, especially suited for applications with limited defect priors and scarce data. Inspired by this, we introduce a support branch based on reference images to emulate the human intuition of anomaly discrimination through contrastive understanding. For each category's k-shot normal samples, one image is consistently selected as the support set during both training and testing. Specifically, the support image I_s is fed into the CLIP-V to extract intermediate feature representations at layer l. The computation proceeds as follows:

$$F_s^{(l)} = \text{CLIP-V}^{(l)}\left(I_s\right) \in \mathbb{R}^{B\times C\times H\times W} \tag{8}$$

Although the support set and query set belong to the same product, they may still exhibit significant spatial discrepancies due to differences in acquisition viewpoint, pose variations, and scale changes. For example, in the MVTec AD [43], categories such as metal_nut and screw exhibit high sensitivity to rotation, making their visual features difficult to spatially align precisely. Under such conditions, directly computing feature similarity between the two may lead to semantic misalignment, thereby compromising accurate modeling and discrimination of anomaly regions. To address this issue, we introduce an Adapter before comparing the features of query and support images, aiming to structurally adapt the query set features and improve their robustness to spatial differences. The Adapter employs the identical architecture as the Adapter$_t$ described in Section III-C. Specifically, the query image feature $F_q^{(l)}$ extracted from the l^{th} layer is input into the Adapter$_s^{(l)}$ yielding the enhanced representation $\tilde{F}_q^{(l)}$, which is then compared pixel-wise with the support image feature $F_s^{(l)}$ to compute the perceptual contrast map $S^{(l)}$ for the support branch, as calculated below:

$$\tilde{F}_q^{(l)} = \text{Adapter}_s^{(l)}\left(F_q^{(l)}\right) \in \mathbb{R}^{B\times C\times H\times W} \tag{9}$$

$$\begin{aligned} S^{(l)} &= \sigma\left(1 - \text{Sim}\left(\tilde{F}_q^{(l)}, F_s^{(l)}\right)\right) \in \mathbb{R}^{B\times 1\times H\times W} \\ &= \sigma\left(1 - \frac{\tilde{F}_q^{(l)} \cdot F_s^{(l)}}{\|\tilde{F}_q^{(l)}\|_2 \, \|F_s^{(l)}\|_2}\right) \end{aligned} \tag{10}$$

Where $\tilde{F}_q^{(l)}$ denotes the query features refined for spatial discrepancy adaptation, and $S^{(l)}$ corresponds to the perceptual contrast map, which encodes the spatial discrepancy at the l^{th} layer between the query and support images.

E. Semantic Fusion and Classifier

To integrate multi-scale semantic similarity information, we compute both language-aware similarity and perceptual contrast maps from multiple layers, concatenate them along the channel dimension, and feed them into a subsequent classifier to produce the final anomaly prediction map, as illustrated below:

$$P = \frac{1}{4}\sum_{l=1}^{4} P^{(l)} \in \mathbb{R}^{B\times 2\times H\times W} \tag{11}$$

$$S = \text{Cat}\left(S^{(1)}, S^{(2)}, S^{(3)}, S^{(4)}\right) \in \mathbb{R}^{B\times 4\times H\times W} \tag{12}$$

$$A = \text{Classifier}\left(\text{Cat}(P, S)\right) \in \mathbb{R}^{B\times 2\times H\times W} \tag{13}$$

Where A represents the final anomaly prediction map, cat$(\cdot)$ denotes the channel-wise concatenation operation, and Classifier$(\cdot)$ denotes a 1×1 convolutional classifier with a softmax activation, which is designed to automatically integrate the multi-level anomaly cues from both language-aware and perceptual-contrast branches and generate two probability maps corresponding to the "normal" and "anomalous" classes, thereby enabling fine-grained pixel-wise anomaly perception.

F. Loss Function

During the training phase, the CLIP-V and CLIP-T are frozen, and only the learnable prompts, Adapter, and Classifier are optimized. To guide the model in learning pixel-level anomaly perception capabilities, we employ the Focal loss and Dice loss to perform end-to-end optimization on the Adapter and Classifier:

$$\begin{aligned} \mathcal{L}_{\text{focal}} = &-\alpha(1-A)^{\gamma}\log(A)\cdot M \\ &-(1-\alpha)A^{\gamma}\log(1-A)\cdot(1-M) \end{aligned} \tag{14}$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum(A\cdot M)+\varepsilon}{\sum A + \sum M + \varepsilon} \tag{15}$$

TABLE I

ANOMALY DETECTION AND LOCALIZATION (IMAGE-AUROC/ PIXEL-AUROC, %) ON THE MVTEC AD DATASET EVALUATED UNDER 2-SHOT, 4-SHOT, AND 8-SHOT SETTINGS. - INDICATES THAT THE RESULT WAS NOT REPORTED IN THE ORIGINAL PAPER. METHODS MARKED WITH * REQUIRE AUXILIARY INDUSTRIAL DATASETS FOR TRAINING. THE BEST RESULTS ARE SHOWN IN BOLD.

Category	k=2					k=4					k=8				
	DifferNet	RegAD*	ADFormer*	InCTRL*	SCF-AD	DifferNet	RegAD*	ADFormer*	InCTRL*	SCF-AD	DifferNet	RegAD*	ADFormer*	InCTRL*	SCF-AD
Carpet	78.4/-	96.5/98.9	99.7/98.7	99.7/-	**100/99.6**	78.6/-	97.9/98.9	99.7/98.7	**100**/-	**100/99.6**	78.5/-	98.5/98.9	99.7/98.8	**100**/-	**100/99.6**
Grid	62.1/-	84.0/77.4	96.8/94.7	97.4/-	**99.9/98.6**	60.5/-	91.2/85.7	97.2/96.2	99.8/-	**99.9/98.7**	78.5/-	91.5/88.7	98.6/97.1	**99.9**/-	**99.9/98.9**
Leather	90.7/-	99.4/98.0	**100**/99.3	**100**/-	**100/99.7**	91.2/-	**100**/99.1	**100**/99.3	**100**/-	**100/99.6**	92.2/-	**100**/98.9	**100**/99.3	**100**/-	**100/99.6**
Tile	97.0/-	94.3/94.3	98.4/94.6	**100**/-	98.6/**98.6**	98.0/-	95.5/94.9	98.7/94.7	**99.9**/-	99.1/**98.8**	99.6/-	97.4/95.2	99.0/94.8	**99.8**/-	99.4/**98.7**
Wood	98.1/-	99.2/93.5	**99.7**/95.0	99.6/-	98.5/**97.2**	96.4/-	98.6/94.7	**99.6**/95.2	99.5/-	98.7/**97.0**	99.4/-	99.4/94.6	**99.8**/95.2	99.6/-	98.4/**97.4**
Bottle	99.3/-	99.4/98.0	**99.8/98.3**	**99.8**/-	99.7/97.8	99.3/-	99.4/**98.4**	99.6/**98.4**	**100**/-	99.9/97.5	99.4/-	99.8/97.5	99.9/**98.4**	**100**/-	**100**/97.8
Cable	85.3/-	65.1/91.7	**95.3/94.3**	88.4/-	84.7/92.1	85.3/-	76.1/92.7	**96.2/94.4**	88.8/-	80.3/92.0	87.9/-	80.6/**94.9**	**97.1**/94.4	93.2/-	83.9/92.0
Capsule	73.0/-	67.5/97.3	78.5/**97.9**	**86.0**/-	85.9/96.9	80.3/-	72.4/97.6	81.9/**98.2**	86.8/-	**94.6/98.2**	78.6/-	76.3/98.2	87.1/**98.5**	89.0/-	**95.3**/98.3
Hazelnut	94.9/-	96.0/98.1	**99.5/98.6**	96.8/-	99.1/97.9	95.8/-	95.8/98.0	**99.6/98.7**	97.2/-	99.3/98.5	97.9/-	96.5/98.5	**99.9/98.9**	97.5/-	99.7/98.3
Metal_nut	61.9/-	91.4/**96.9**	97.8/94.1	94.2/-	**98.7**/85.0	67.3/-	94.6/**97.8**	98.8/94.9	94.6/-	**99.2**/86.6	67.6/-	98.3/96.9	98.9/95.3	94.7/-	**99.9/88.4**
Pill	83.2/-	81.3/**93.6**	91.5/93.4	89.3/-	**94.5**/90.6	84.0/-	80.8/**97.4**	92.4/93.7	87.4/-	**93.9**/90.9	82.1/-	80.6/**97.8**	93.4/93.9	88.8/-	**94.6**/91.2
Screw	73.4/-	52.5/94.4	82.0/**98.4**	80.3/-	**86.2/98.4**	72.5/-	56.6/95.0	85.6/**98.6**	81.5/-	**89.1**/98.2	75.0/-	63.4/97.1	86.3/**98.9**	82.5/-	**90.1**/98.5
Toothbrush	60.8/-	86.6/98.2	85.9/98.1	96.0/-	**97.8/98.6**	62.5/-	90.9/98.5	90.1/98.6	96.3/-	**98.6/98.8**	60.8/-	98.5/98.7	97.8/**99.1**	97.5/-	**99.7**/98.9
Transistor	61.8/-	86.0/**93.4**	**93.6**/87.9	87.5/-	93.4/92.8	62.2/-	85.2/93.8	**95.6**/89.7	89.5/-	94.8/**93.8**	63.3/-	93.4/96.8	**97.5**/90.3	90.3/-	96.2/**95.5**
Zipper	89.2/-	86.3/95.1	89.8/98.6	95.4/-	**99.9/99.0**	84.8/-	88.5/94.0	93.0/98.7	96.4/-	**100/99.1**	87.3/-	94.0/97.4	93.5/98.8	97.0/-	**100/99.3**
Means	80.6/-	85.7/94.6	93.9/96.1	94.0/-	**95.8/96.2**	81.3/-	88.2/95.8	95.2/**96.5**	94.5/-	**96.5/96.5**	83.2/-	91.2/96.7	96.6/**96.8**	95.3/-	**97.1/96.8**

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{focal}} + \lambda \mathcal{L}_{\text{dice}} \tag{16}$$

Where A denotes the anomaly probability map output by the model, and M represents the pixel-level pseudo labels generated by the synthetic anomaly module. The parameters α, γ, ε, λ are set to 1, 2, 1×10^{-5} and 1, respectively.

IV. EXPERIMENTS

To validate the effectiveness of SCF-AD, systematic comparative experiments are conducted on four challenging benchmarks: MVTec AD, VisA, BTAD, and MPDD. Furthermore, detailed ablation studies on the MVTec AD dataset and a real-world test on battery spot-welding defects are performed to assess its practical applicability.

A. Experimental Setting

1) **Dataset**: We conducted comparative experiments on the MVTec AD [43], VisA [44], BTAD [52], and MPDD [53] datasets. All ablation studies were performed on the MVTec AD dataset, which covers 15 industrial product categories with diverse defect types and is widely used to assess the accuracy and robustness of industrial anomaly detection methods. The VisA dataset is a more recent and challenging benchmark with 12 categories featuring complex scenes and subtle defects. The BTAD dataset provides 3 additional distinct industrial product categories for evaluation. The MPDD dataset includes 6 categories of metal parts with significant spatial and viewpoint variations. During training, pseudo anomalies were synthesized using texture backgrounds sampled from the Describable Textures Dataset [47]. To evaluate real-world applicability, we further constructed a battery spot-welding anomaly detection dataset (BSW AD [54], [55]), as shown in Fig. 6, consisting of 8,184 images captured by AOI systems across 10 typical welding types. Among them, 922 images are pixel-wise annotated with defects such as Welding Burn, Welding Shift, and Missing Welds. The dataset features complex backgrounds and subtle structural variations, reflecting the practical challenges of industrial inspection tasks.

Evaluation Metrics: We conducted a comprehensive evaluation using six standard metrics. At the image level, we assessed anomaly detection performance using the Area Under the Receiver Operating Characteristic curve (Image-AUROC), the Area Under the Precision-Recall curve (Image-AUPR), and the maximum F1-score (Image-F1-max). At the pixel level, to accurately measure anomaly localization quality, we employed the pixel-wise AUROC (Pixel-AUROC), the maximum pixel-wise F1-score (Pixel-F1-max), and the Per-Region Overlap (Pixel-PRO), which is specifically designed to evaluate the overlap coverage of varying anomaly regions.

B. Comparison with State-of-the-art Methods

To validate the performance of SCF-AD, we conducted comparisons on the MVTec AD dataset, the VisA dataset, BTAD dataset, and the MPDD dataset against several representative methods, including DifferNet [56], PatchCore [13], RegAD [20], ADFormer [21], InCTRL [32], WinCLIP [41], APRIL-GAN [31], FADE [30], AnomalyGPT [42], and INP-Former [57]. It is worth noting that RegAD and ADFormer adopt a "leave-one-out" training strategy, in which 14 out of the 15 categories in MVTec are used as the training set. InCTRL, APRIL-GAN, and AnomalyGPT are trained on the VisA and tested on the target classes of MVTec AD, requiring k-shot samples of the target class as references. In contrast, our method relies solely on the k-shot normal images of the target

TABLE II

ANOMALY DETECTION AND LOCALIZATION (%) ON THE MVTEC AD, VISA, AND BTAD DATASETS EVALUATED UNDER 1-SHOT, 2-SHOT, AND 4-SHOT SETTINGS. METHODS MARKED WITH * REQUIRE AUXILIARY INDUSTRIAL DATASETS FOR TRAINING. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Dataset	Method	Image-AUROC			Image-AUPR			Image-F1-MAX			Pixel-AUROC			Pixel-PRO			Pixel-F1-MAX		
		k=1	k=2	k=4	k=1	k=2	k=4	k=1	k=2	k=4	k=1	k=2	k=4	k=1	k=2	k=4	k=1	k=2	k=4
MVTec	PatchCore	83.4	86.3	88.8	92.2	93.8	94.5	90.5	92.0	92.6	92.0	93.3	94.3	82.3	79.7	84.3	53.0	50.4	55.0
	WinCLIP	93.1	94.4	95.2	96.5	97.0	97.3	93.7	94.4	94.7	95.2	96.0	96.2	87.1	88.4	89.0	55.9	58.4	59.5
	AnomalyGPT*	94.1	95.5	96.3	–	–	–	–	–	–	95.3	95.6	96.2	–	–	–	–	–	–
	FADE*	93.9	95.2	96.3	96.8	97.6	98.1	94.8	95.0	95.5	95.4	95.8	96.2	88.3	88.9	89.5	54.6	55.8	57.0
	APRIL-GAN*	92.4	92.6	92.8	–	–	–	92.4	92.6	96.0	95.1	95.5	95.9	90.6	91.3	91.8	54.2	55.9	56.9
	INP-Former	**96.6**	**97.0**	**97.6**	**98.2**	98.2	**98.6**	**96.4**	**96.7**	**97.0**	**97.0**	**97.2**	**97.0**	**92.6**	**93.1**	**92.9**	**64.0**	**65.6**	**65.6**
	SCF-AD(Ours)	95.0	95.8	96.5	98.0	**98.3**	98.5	95.2	95.7	96.2	95.0	96.2	96.5	90.1	91.3	92.2	53.0	56.2	57.2
VisA	PatchCore	79.9	81.6	85.3	–	–	–	81.7	82.5	84.3	95.4	96.1	96.8	80.5	82.6	**94.9**	38.0	41.0	43.9
	WinCLIP	83.8	84.6	87.3	85.1	85.8	88.8	83.1	83.0	84.2	96.4	96.8	97.2	85.1	86.2	87.6	41.3	43.5	47.0
	AnomalyGPT*	87.4	88.6	90.6	–	–	–	–	–	–	96.2	96.4	96.7	–	–	–	–	–	–
	FADE*	86.7	89.2	91.9	87.9	90.2	91.9	84.7	85.9	87.0	**97.5**	**97.8**	**98.0**	88.9	89.8	90.0	42.3	44.4	46.4
	APRIL-GAN*	91.2	92.2	92.6	–	–	–	86.9	87.7	88.4	96.0	96.2	96.2	90.0	90.1	90.2	38.5	39.3	40.0
	INP-Former	**91.4**	**94.6**	**96.4**	92.2	**94.9**	**96.0**	**91.8**	**90.8**	**93.0**	96.3	97.2	97.7	89.5	**91.8**	93.1	**47.3**	**50.4**	**54.3**
	SCF-AD(Ours)	**91.4**	92.3	92.6	**92.7**	93.7	94.0	88.0	88.5	89.3	96.7	97.0	97.2	86.4	87.2	88.9	37.1	38.1	38.2
BTAD	PatchCore	88.8	92.8	94.1	86.6	93.7	95.8	82.8	88.2	**92.7**	96.0	96.7	97.0	74.9	76.6	78.2	45.5	48.7	50.8
	WinCLIP	84.9	85.9	87.0	79.7	82.2	83.8	77.8	78.8	80.1	95.6	96.5	95.8	66.0	66.0	66.7	47.9	49.2	49.6
	AnomalyGPT*	65.1	64.7	64.7	73.2	73.3	73.2	79.6	80.6	80.9	78.5	78.6	78.4	70.3	67.0	66.4	22.8	23.2	22.6
	FADE*	69.7	73.0	75.2	71.4	72.5	72.9	71.0	70.0	73.7	85.5	86.6	88.0	50.7	52.4	54.9	29.9	30.7	33.5
	APRIL-GAN*	92.5	92.9	92.8	**95.7**	**96.4**	95.3	**91.8**	91.8	91.8	94.0	94.4	94.4	79.4	79.2	79.4	52.5	53.9	54.0
	INP-Former	89.6	91.8	93.2	96.5	97.2	**97.2**	**91.8**	92.2	92.3	96.4	**97.3**	**97.2**	73.1	77.3	78.1	**63.7**	**65.1**	**65.6**
	SCF-AD(Ours)	**93.7**	**94.0**	**94.2**	95.3	95.9	96.1	91.6	**92.3**	**92.7**	**96.8**	96.8	97.0	77.1	78.9	78.4	57.8	60.5	59.3
MPDD	PatchCore	63.1	62.0	67.1	67.2	64.7	74.5	79.4	79.0	83.7	95.8	96.3	96.9	88.2	89.3	90.2	24.1	24.3	30.7
	WinCLIP	70.7	73.0	75.5	72.9	75.3	80.1	81.3	82.1	83.7	96.3	96.5	94.8	88.5	89.9	89.1	31.5	33.5	35.1
	AnomalyGPT*	75.7	74.0	73.0	77.4	75.8	73.6	82.2	81.3	80.9	**96.6**	96.4	96.7	78.6	77.7	77.2	34.8	34.1	33.0
	FADE*	63.2	68.2	68.1	71.9	75.3	76.5	79.7	82.0	81.0	85.8	88.9	89.1	68.6	71.1	71.3	16.1	17.1	18.4
	APRIL-GAN*	79.9	82.2	85.6	83.8	86.8	90.0	81.9	84.9	86.4	**96.6**	96.6	96.7	**90.9**	91.1	91.0	**40.1**	**40.5**	41.1
	INP-Former	74.5	78.3	82.3	79.3	82.7	83.5	81.5	83.3	88.1	96.2	**97.3**	**97.7**	89.4	**92.8**	**93.6**	32.0	40.2	40.0
	SCF-AD(Ours)	**85.3**	**88.8**	**90.6**	**86.1**	**91.1**	**92.6**	**85.1**	**90.2**	**91.1**	96.3	97.0	97.3	88.9	89.5	91.3	35.5	36.1	**41.4**

category, without requiring any auxiliary industrial data. The evaluation results of the aforementioned methods are taken from their original papers. Unless otherwise specified, all SCF-AD results are averaged over three independent runs under identical settings.

To ensure an accurate and fair comparison, we adopt the state-of-the-art results directly from the original publications of each method, unless otherwise specified. Similarly, some evaluation metrics are marked with "-" as these results were not available in the corresponding papers. A key exception is the BTAD and the MPDD datasets, for which public few-shot baseline results are not available. For these benchmarks, we re-implemented the competing baseline methods to provide a comprehensive and fair comparison. Since these studies employ different few-shot evaluation settings (e.g., k=1, 2, 4 and k=2, 4, 8), we present our comparative analysis in separate tables to align with their respective frameworks. We evaluated our proposed SCF-AD under all of these k-shot settings to guarantee a direct and fair comparison against each group of methods.

Comparison on the MVTec AD dataset. Tables I and II present the performance of state-of-the-art anomaly detection methods on the MVTec dataset. In Table I, our method consistently achieves the highest mean scores in both image-level and pixel-level AUROC across all settings (k=2, k=4, and k=8). In Table II, facing the latest competitors, SCF-AD maintains highly competitive results. While INP-Former shows strong performance, our method ranks within the top two in the majority of settings. For instance, in terms of Image-AUROC, SCF-AD consistently follows INP-Former as the second-best method across $k = 1, 2,$ and 4. In pixel-level localization, SCF-AD also secures strong performance, ranking among the top methods and surpassing APRIL-GAN and FADE in Pixel-AUROC at $k = 2$ and $k = 4$. This demonstrates that SCF-AD remains a robust framework for industrial anomaly detection. Qualitatively, Fig. 3(e) illustrates the visual detection results of SCF-AD on the MVTec dataset, demonstrating its excellent capability in localizing anomaly regions.

Comparison on the VisA dataset. Table II summarizes the performance on the VisA dataset. SCF-AD continues to demonstrate excellent generalization ability, securing the second-best results in Image-AUROC, Image-AUPR, and Image-F1-MAX across the majority of k-shot settings. Notably, at $k = 1$, SCF-AD matches the performance of INP-Former in Image-AUROC of 91.4% and achieves the highest Image-AUPR of 92.7%, proving its effectiveness even in extremely low-shot scenarios. Qualitatively, the visualization results in Fig. 4 further confirm SCF-AD's robustness on the VisA dataset, showcasing its ability to precisely localize defects across diverse categories.

Comparison of the BTAD dataset. Table II presents the image-level and pixel-level detection performance of the competing methods on the BTAD dataset. Our method demonstrates a strong and robust performance, consistently

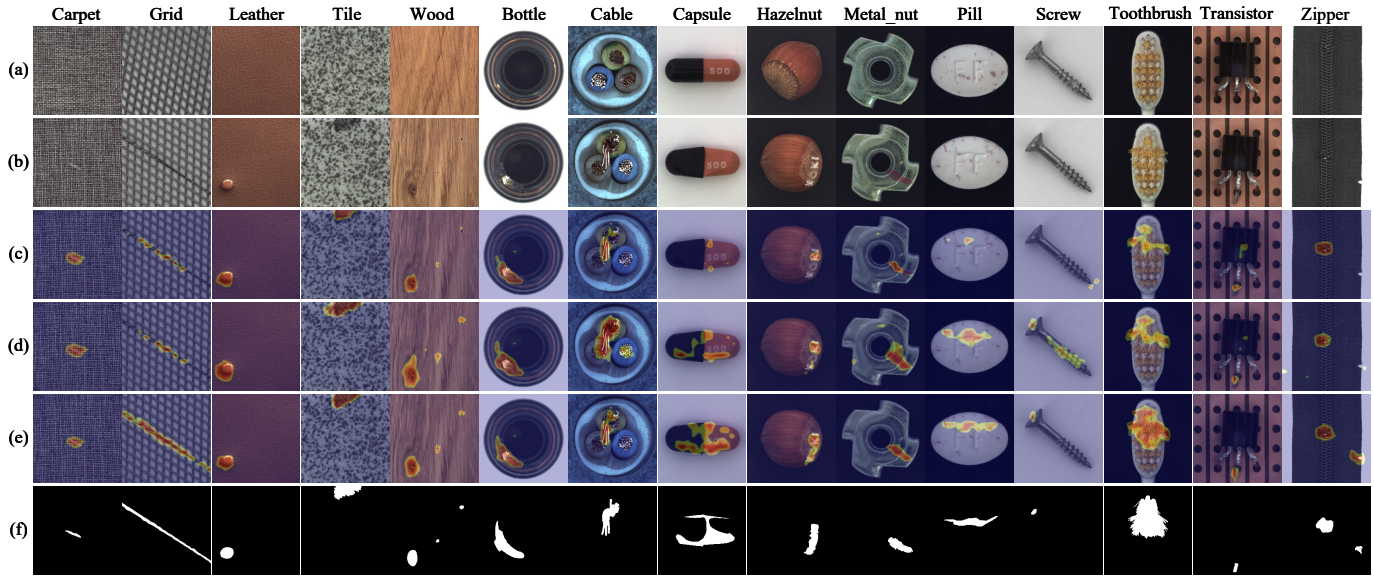

Fig. 3. Qualitative visualization results on the MVTec AD dataset. (a) Normal sample. (b) Anomaly sample. (c) w/o Support Set. (d) w/o Adapter. (e) Ours. (f) Ground Truth.

achieving the best results in Image-AUROC across all k-shot settings (k=1, 2, and 4). Regarding pixel-level localization, our method exhibits exceptional precision, particularly in the 1-shot setting. Notably, at $k = 1$, SCF-AD achieves a leading Pixel-AUROC of 96.8%, surpassing INP-Former. Furthermore, while slightly trailing INP-Former at $k = 2$ and $k = 4$, our method consistently ranks as the second-best. Qualitatively, the visualization results in Fig. 4 further validate the efficacy of SCF-AD on the BTAD dataset.

Comparison on the MPDD dataset. Table II presents the image-level and pixel-level detection performance of the competing methods on the MPDD dataset. SCF-AD demonstrates clear dominance in image-level detection, achieving the best results in Image-AUROC, Image-AUPR, and Image-F1-MAX across all k-shot settings. Regarding pixel-level localization, our method consistently ranks as the second-best in Pixel-AUROC across all k-shot settings. Qualitatively, the visualization results in Fig. 4 further confirm the robustness of SCF-AD on the MPDD dataset, showcasing its precise anomaly localization under viewpoint transformations.

TABLE III
ABLATION STUDY ON KEY COMPONENTS (IMAGE-AUROC/ PIXEL-AUROC, %)

ID	Support Set	Learnable Token	Adapter	k=2	k=4	k=8
A	✓	✓	×	93.8/93.2	94.5/92.3	96.0/93.8
B	✓	×	✓	74.8/81.7	67.7/76.5	67.5/78.2
C	×	✓	✓	95.5/95.1	96.2/95.6	96.6/95.8
D	✓	×	×	67.7/77.4	72.2/79.1	74.5/77.0
E	×	×	✓	94.2/95.2	95.7/95.5	95.8/95.6
F	×	✓	×	95.0/95.3	95.8/95.5	95.8/95.4
G	✓	✓	✓	**96.0/96.2**	**96.6/96.6**	**97.2/96.8**

TABLE IV
ABLATION STUDY ON PSEUDO-ANOMALY GENERATION RATIOS (IMAGE-AUROC/ PIXEL-AUROC, %)

ID	α (Normal)	β (Texture)	γ (Structure)	k=2	k=4	k=8
A	0.3	0.7	0.0	95.5/95.5	96.8/96.1	96.5/96.0
B	0.3	0.0	0.7	94.6/95.6	95.2/95.8	96.1/96.1
C	0.3	0.5	0.2	**95.8/96.2**	**96.5/96.5**	**97.1/96.8**
D	0.3	0.2	0.5	95.1/95.5	95.9/96.3	96.0/96.3
E	0.3	0.35	0.35	94.7/95.6	96.2/96.1	96.5/96.2
F	0.0	1.0	0.0	72.3/87.5	72.6/87.4	74.7/87.3
G	0.0	0.0	1.0	93.8/95.1	93.4/95.4	95.9/95.9
H	1.0	0.0	0.0	95.7/95.4	96.2/95.9	96.2/96.2

C. Ablation Study

We conducted ablation experiments on the MVTec AD dataset under few-shot settings with k=2, k=4, and k=8 to evaluate the model's performance and analyze the impact of key components.

1) **The Influence of Adapter**: Quantitatively, we designed three groups of comparative experiments (Table III) to evaluate the effectiveness of the Adapter and analyze its performance variations under different component configurations. First, under the full configuration (A vs. G), introducing the Adapter consistently improved performance across all k-shot settings. For example, under the k=8 setting, Image-AUROC improved from 96.0% to 97.2% and Pixel-AUROC from 93.8% to 96.8%. Similar gains were observed under the k=2 and k=4 conditions, confirming that the Adapter effectively mitigates cross-modal discrepancies and enhances structural perception. However, without Learnable Tokens (D vs. B), the Adapter did not yield consistent performance gains, and even showed a significant drop under the k=4 and k=8 settings. The Adapter's adaptation ability depends on a stable semantic reference signal. Without learnable tokens, the model lacks a clear semantic direction, making it difficult for the Adapter to achieve precise alignment. This may introduce errors and interfere with structural modeling, leading to misjudgments. Moreover, even without using the support set (F vs. C), the introduction of the Adapter still led to performance improvements.

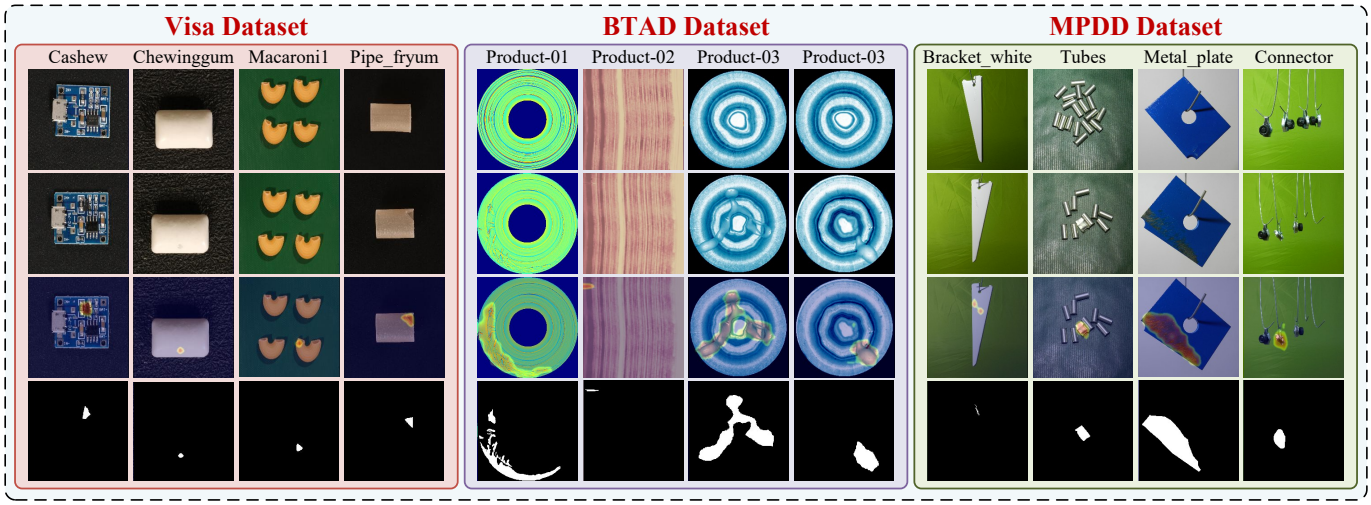

Fig. 4. Qualitative visualization results on the Visa, BTAD, and MPDD datasets. From top to bottom, the rows represent normal samples, defective samples, detection results, and the ground truth.

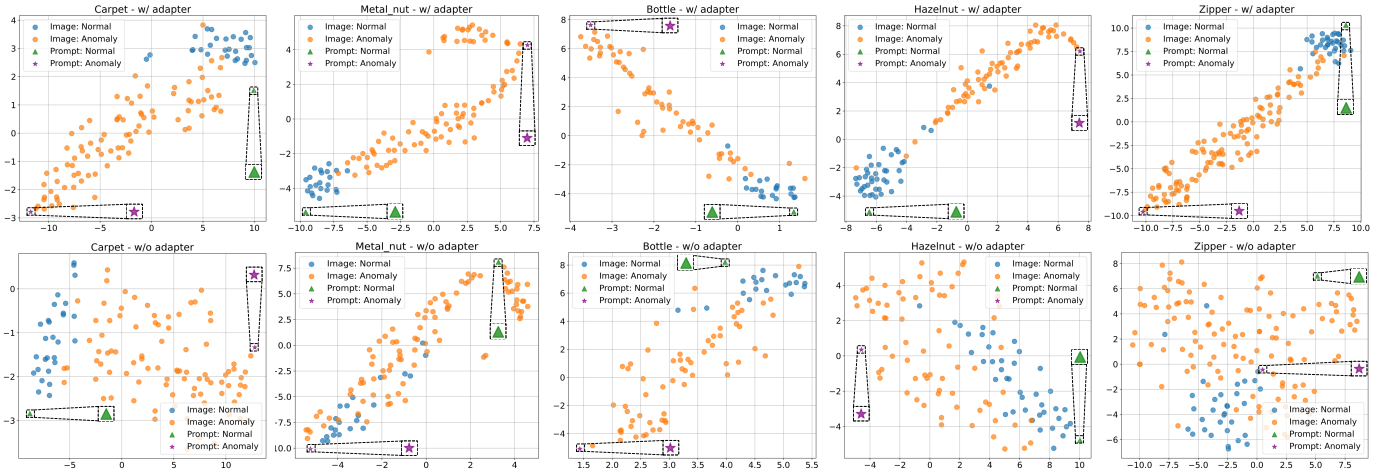

Fig. 5. Feature Distribution Visualization using t-SNE with and without Adapter. Normal image and normal prompt features are compactly aligned, while anomaly prompt features are more dispersed and further away from the normal image features in the embedding space with Adapter.

To evaluate whether the Adapter can effectively mitigate the modality shift issue of CLIP in industrial semantic scenarios, we conducted a qualitative analysis on the alignment between query image features and prompt text features. Specifically, test set from five categories and associated prompts were embedded and visualized via t-SNE in 2D semantic space (Fig. 8). The results show that after introducing the Adapter, the features of normal images cluster closely with the "Normal" prompt, while the "Anomaly" prompt is clearly separated, leading to sharper semantic boundaries. In contrast, without the Adapter, the distributions of image and text features are disordered, especially in categories such as carpet, hazelnut, and zipper, showing poor modal alignment and weakened semantic distinction between normal and anomalous states. Moreover, to intuitively demonstrate the impact of Adapter on localization accuracy, Fig. 3(d) and (e) illustrate the results without and with Adapter, respectively. It can be observed that the model achieves more accurate localization of anomalous regions after incorporating the Adapter (Fig. 3(e)).

To verify the robustness of the Adapter against spatial structural disturbances, we designed a rotation robustness test experiment. As shown in Fig. 7, we selected the metal_nut and screw categories from the MVTec AD dataset, which exhibit no orientation variations across samples, and applied random rotations (0° to ± 30°) to the test set to evaluate the variation in model performance under different angles. In the experiments, we compared the performance with and without the Adapter under k = 2, 4, and 8 settings, and averaged the results over three random seeds. The experimental results show that the model performance degrades to some extent after rotation, mainly due to the interpolation artifacts introduced during rotation, which create boundary regions inconsistent with the background and lead to misclassification. However, compared with models without the Adapter module, those equipped with Adapter exhibit significantly reduced performance fluctuations, indicating improved robustness.

2) **The Influence of Learnable Token**: To evaluate the effectiveness of the Learnable Token, we conducted three sets

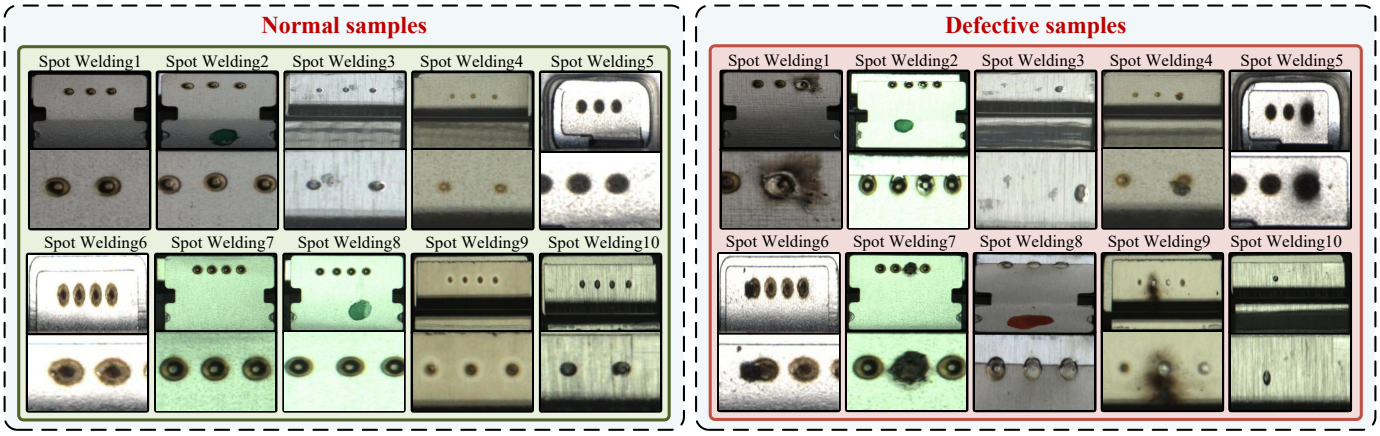

Fig. 6. Samples from the BSW AD dataset. The left panel contains normal samples and the right panel contains defective samples. Each category consists of two rows: the first row is the original image, and the second is a magnified close-up view.

of comparative experiments under different module configurations (Table III). In the complete configuration (G vs B, k=2), the introduction of the Learnable Token improved Image-AUROC and Pixel-AUROC by 21.2% and 14.5%, respectively, indicating its significant contribution to aligning with semantic prompts. Even without the support set branch (C vs E) or the Adapter (A vs D), the Learnable Token still delivered noticeable performance improvements, further confirming its ability to provide stable semantic guidance.

3) **The Influence of Support Set**: To evaluate its effectiveness, we compared the performance changes with and without this component under different configurations (Table III). When lacking the Adapter or semantic guidance (F vs A, E vs B), introducing the support set actually led to performance degradation. Although the support set provides normal sample references, without Learnable Tokens, the model lacks a semantic anchor, making the structural comparison semantically ambiguous; without an alignment mechanism, the structural references are prone to mismatches and false positives. Moreover, to intuitively demonstrate the impact of Adapter on localization accuracy, Fig. 3(c) and (e) illustrate the results without and with support set, respectively. It can be observed that the model achieves more accurate localization of anomalous regions after incorporating the Adapter (Fig. 3(e)).

4) **The Influence of Anomaly Generation**: To evaluate how the composition ratio of anomalous samples affects model performance, we define the proportions of normal images, texture-based and structure-based pseudo anomalies in the training set as α, β, and γ, respectively, where $\alpha + \beta + \gamma = 1$. This composition strategy not only determines the distribution of anomalous samples in training but also directly influences the model's attention to and perception of different anomaly types. For instance, in MVTec AD, categories like leather, grid, and wood mainly exhibit texture anomalies, while transistor and metal_nut are dominated by structural defects (e.g., misplaced and bending), requiring stronger structural modeling. Experimental results show that a pseudo-anomaly composition of "0.3 normal / 0.5 texture / 0.2 structure" consistently yields the best performance across various few-shot settings (Table IV), confirming the generalization capability of this design

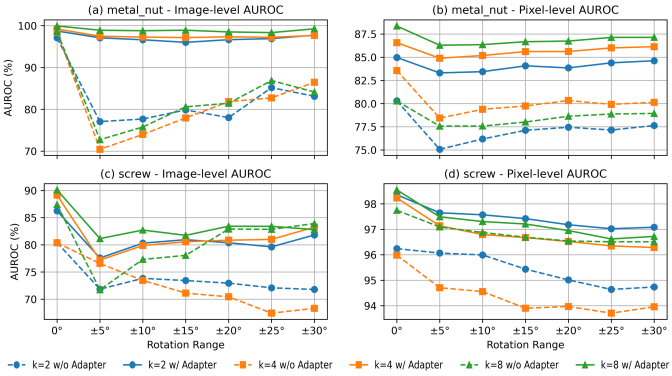

Fig. 7. Robustness analysis against spatial misalignment with and without Adapter. Models equipped with Adapter exhibit consistently more stable performance across varying rotation degrees, demonstrating improved robustness to spatial misalignment.

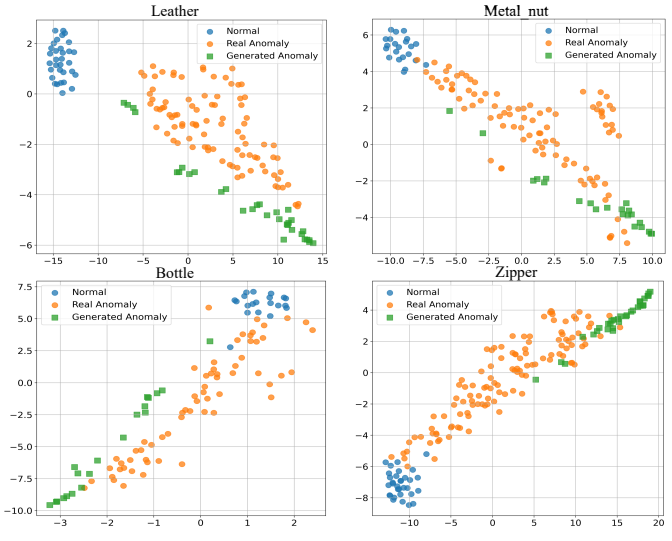

Fig. 8. Feature Distribution Visualization using t-SNE for Normal (blue), Real Anomaly (orange), and Generated Anomaly (green) Samples.

across diverse industrial anomaly patterns.

In addition, the t-SNE visualizations in Fig. 8 provide

further insight into the learned feature space. Across all representative categories shown (Leather, Metal_nut, Bottle, and Zipper), it is clear that our Generated Anomaly samples (green) are almost not overlapped with the Real Anomaly samples (orange). Despite this distributional gap, the learned representation is clearly able to distinguish between the compact Normal cluster (blue) and both types of anomalies. This strongly suggests that while pseudo-anomaly generation is not a perfect simulation of real defects, learning to identify these synthetic irregularities generalizes well to detecting unseen, real anomalies.

5) **The Influence of Feature Matching Strategies**: To assess the effect of different feature matching strategies in the human perceptual contrast matching branch, we compare L1, L2, and cosine similarity under k=2, 4, and 8 few-shot settings (Table V). Experimental results show that cosine similarity consistently achieves the best performance. For instance, under the k=2 condition, it improves Image-AUROC by 0.7% over L2-distance, and Pixel-AUROC by 2.1% over L1-distance, demonstrating its comprehensive advantage. These results indicate that cosine similarity is more effective in measuring the semantic similarity between support and query images. In contrast, Euclidean distance is heavily influenced by feature magnitudes, making it difficult to precisely characterize directional differences in semantic space and prone to overfitting to low-level details. Cosine similarity focuses on directional consistency of features, enabling more effective alignment of structural semantic features in high-dimensional spaces. Notably, in MVTec AD categories such as metal_nut and screw, even normal samples exhibit significant rotational variations. Such samples, though semantically similar, may produce large feature deviations when using L1 or L2-distance, leading to misclassification. Cosine similarity, however, is robust to magnitude shifts caused by such rotational perturbations.

6) **The Influence of Support Set Selection Strategies**: To systematically evaluate how the selection strategy of the support set affects model performance, we compared three approaches: Fixed One (selecting a single image from the training set), and feature fusion strategies based on Mean and PCA over four scales of CLIP-Vision features. As shown in Table VI, under the k=2 setting, the Pixel-AUROC improves from 95.5% to 96.2% and the Image-AUROC from 95.2% to 95.8% when using Fixed One instead of the Mean strategy. Similar improvements are observed for both metrics under the k=4 and k=8 settings. The support set selected by the Fixed One strategy preserves complete spatial structures and semantic context, avoiding the loss of critical regional features that may occur during mean fusion or PCA, providing a more discriminative structural reference during the matching phase.

7) **The Influence of Different Backbones**: To evaluate the impact of backbone networks on model performance, we compare five CLIP-pretrained visual encoders: the ResNet series RN50 and RN101, and the Vision Transformer architectures ViT-B/16, ViT-B/32, and ViT-L/14 (Table VII). Higher-resolution Vision Transformers (e.g., ViT-B/16 and ViT-L/14) consistently outperform ResNet counterparts, with larger models delivering further gains. Under the extremely limited k=2 setting, ViT-B/16 achieves 92.8% Image-AUROC and 95.1% Pixel-AUROC, surpassing RN50 (89.0% / 93.7%) by a clear margin. ViT-L/14 delivers the best performance across all settings, reaching 97.1% Image-AUROC and 96.8% Pixel-AUROC at k=8, indicating superior generalization ability and enhanced anomaly localization. This improvement stems from the strong representation capability of ViT, which, unlike the convolutional operations in conventional CNNs, employs a self-attention mechanism to capture long-range dependencies and global context while preserving local details essential for accurate few-shot anomaly detection. In contrast, ViT-B/32 underperforms even the ResNet variants, primarily due to its large patch size (32 × 32), which halves the spatial resolution of feature maps compared to ViT-B/16 and causes loss of fine-grained details critical for detecting small anomalies in industrial images. These results suggest that preserving higher spatial granularity is crucial for achieving high pixel-level localization accuracy in industrial scenarios.

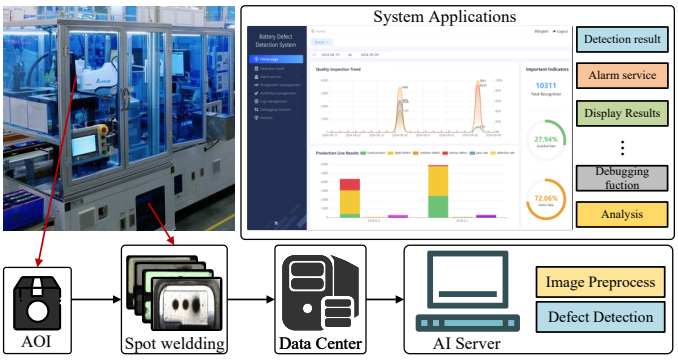

Fig. 9. The practical deployment workflow in a production line, from image acquisition by an AOI device to analysis and final display on the system interface.

TABLE V
ABLATION STUDY ON FEATURE MATCHING STRATEGIES (IMAGE-AUROC/ PIXEL-AUROC, %)

Method	k=2	k=4	k=8
L1-Distance	94.7/94.1	95.7/93.9	96.3/94.2
L2-Distance	95.1/94.1	96.0/94.6	96.3/94.4
Cosine Similarity	**95.8/96.2**	**96.5/96.5**	**97.1/96.8**

TABLE VI
ABLATION STUDY ON SUPPORT SET SELECTION STRATEGIES (IMAGE-AUROC/ PIXEL-AUROC, %)

Method	k=2	k=4	k=8
Mean	95.2/95.5	95.9/95.6	96.4/95.8
PCA	95.6/95.8	95.4/95.7	96.4/96.0
Fixed One	**95.8/96.2**	**96.5/96.5**	**97.1/96.8**

TABLE VII
ABLATION STUDY ON DIFFERENT BACKBONE (IMAGE-AUROC/ PIXEL-AUROC, %)

Backbone	k=2	k=4	k=8
RN50	89.0/93.7	89.4/94.2	90.3/94.6
RN101	88.5/94.0	89.7/94.5	90.2/94.0
ViT-B/16	92.8/95.1	95.8/95.9	95.6/95.9
ViT-B/32	82.2/92.3	85.6/93.9	88.8/94.1
ViT-L/14	**95.8/96.2**	**96.5/96.5**	**97.1/96.8**

D. Real-World Application

In modern manufacturing, the rapid iteration of product models renders it impractical to collect a large number of normal samples for every new model. Consequently, the self-supervised FSAD paradigm is particularly well-suited for such scenarios. To comprehensively evaluate the practical applicability and deployment adaptability of the proposed SCF-AD method in this context, we conducted an application verification on a real-world battery spot-welding task. To this end, we constructed a BSW AD dataset, in which all images were collected from an actual production line using AOI equipment. As shown in Fig. 6, the dataset covers ten representative types of spot welds, including both normal and defective samples, such as Welding Burn, Welding Shift, and Missing Welds. As illustrated in Fig. 9, the operational workflow of the inspection system is as follows: the AOI device captures images of the weld surfaces and transmits them to a data center; subsequently, an AI server equipped with our algorithm performs automatic defect detection; finally, the results are visualized on a front-end interface for real-time monitoring and analysis by plant personnel.

Quantitatively, we conducted a comprehensive comparison of SCF-AD against several state-of-the-art few-shot anomaly detection methods, including WinCLIP, APRIL-GAN, RegAD, and ADFormer, on the BSW AD dataset in terms of both performance and complexity, as shown in Fig. 10. In terms of performance, SCF-AD consistently outperformed all comparison methods across all settings from 1-shot to 8-shot in both Image AUROC and Pixel AUROC. This demonstrates that our method can achieve accurate defect detection and localization even with very few normal samples. Regarding model complexity, although we employed CLIP ViT-L/14 as the backbone with a total parameter size of 432.6M, the GPU memory usage remains only 2,854 MiB, which is significantly lower than most compared methods. Most importantly, SCF-AD achieves a high inference speed of 126.3 FPS, which surpasses all other methods and highlights its well-optimized structure and superior computational efficiency. This high-speed inference capability is particularly important for real-world deployment. Moreover, the total training time was approximately 0.4 hours, demonstrating its potential for rapid model deployment and iteration.

Qualitatively, Fig. 11 presents the visualization results of different methods on various defective samples. The anomaly heatmaps generated by our method align more closely with the actual defect contours and exhibit less background noise, demonstrating superior localization capability. For instance, in the second row's "Missing Welds" sample, only our method successfully and precisely localized the missing welds, while all other methods failed to effectively detect this anomaly.

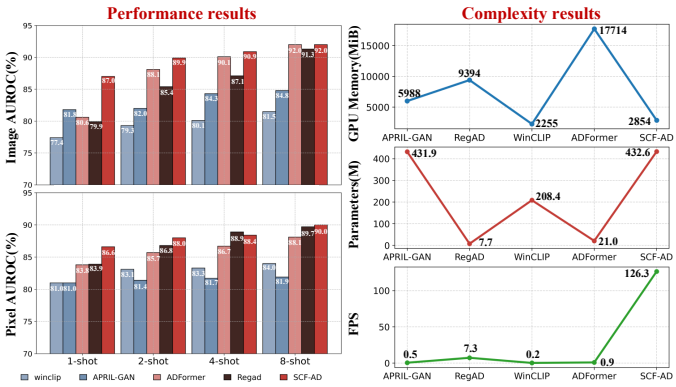

Fig. 10. Quantitative comparison of SCF-AD against other state-of-the-art methods, highlighting its advantages in detection accuracy and model complexity.

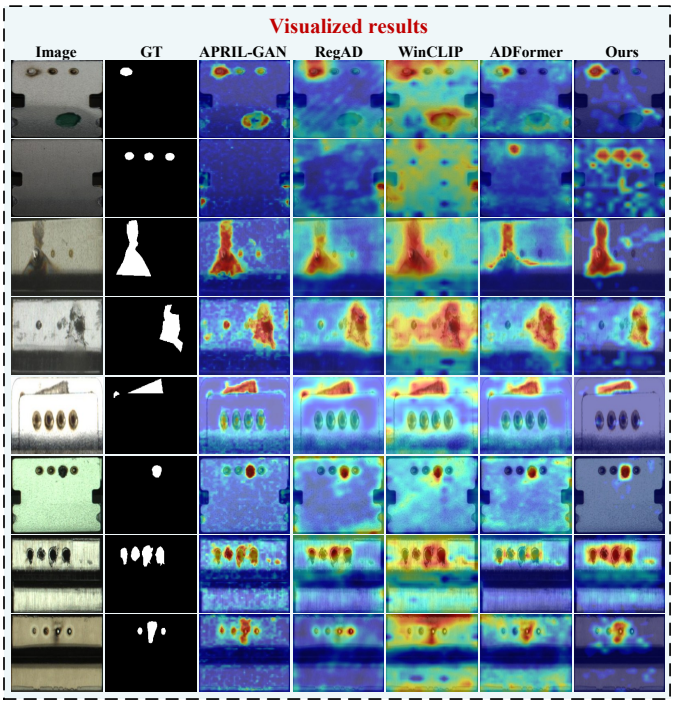

Fig. 11. Qualitative visualization of results on BSW AD dataset.

V. CONCLUSION

This paper addresses the common reliance of existing FSAD methods on external auxiliary industrial data for training, and proposes a novel self supervised CLIP-guided framework for few shot industrial anomaly detection. The proposed approach requires no external auxiliary industrial data and performs anomaly detection and localization using only a few normal samples from the target category. Furthermore, a learnable Adapter is introduced to mitigate the semantic misalignment between text and image modalities in CLIP under industrial scenarios, and to enhance the model's robustness to spatial structural discrepancies between the query and support sets. Extensive experiments on the MVTec AD dataset validate the effectiveness and superiority of the proposed SCF-AD method. In a real-world battery spot-welding defect detection task, our method achieves excellent detection accuracy and inference efficiency, demonstrating its feasibility for practical industrial deployment.

Despite the excellent performance demonstrated by the proposed SCF-AD method, several inherent bottlenecks remain. The current framework still follows a "one-model-per-category" paradigm without a unified detection scheme

across heterogeneous product types, which may restrict its scalability in complex multi-category industrial environments. The support set currently relies on a single reference image, which limits the exploitation of few-shot diversity. A memory-based aggregation of all support embeddings may provide a more stable prototype for adaptive query matching. Moreover, in the Language-Aware Visual Alignment branch, the uniform averaging of prompt embeddings blurs the semantic hierarchy and weakens discriminative cues. Although the pseudo-anomaly generation strategy effectively enhances model perception, a semantic gap still exists between synthesized and real defects. Future work will therefore focus on addressing these bottlenecks through unified multi-category architectures, representative prompt learning, and high-fidelity anomaly synthesis (e.g., diffusion-based generation) to further improve model generalization and deployment efficiency.

REFERENCES

[1] K. Lei and Z. Qi, "A dual-state-based surface anomaly detection model for rail transit trains using vision-language model," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–13, 2025.

[2] Y. Zhao, X. Wu, and J. Cheng, "Dualflow: Dual-branch flow for unsupervised anomaly detection and localization," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–14, 2025.

[3] H. Yao, Y. Cao, W. Luo, W. Zhang, W. Yu, and W. Shen, "Prior normality prompt transformer for multiclass industrial image anomaly detection," *IEEE Trans. Ind. Inform.*, vol. 20, no. 10, pp. 11 866–11 876, Oct. 2024.

[4] A. H. Madessa, J. Dong, E. Rigall, Q. Lv, H. S. N. Nawaz, I. Mugunga, and S. Guo, "Transmittance surface detection and material identification using multitask vit-sift fusion," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–17, 2023.

[5] G. C. Giakos, L. Fraiwan, N. Patnekar, S. Sumrain, G. B. Mertzios, and S. Periyathamby, "A sensitive optical polarimetric imaging technique for surface defects detection of aircraft turbine engines," *IEEE Trans. Instrum. Meas.*, vol. 53, no. 1, pp. 216–222, 2004.

[6] Y. Xu, J. Chen, Y. Liang, Y. Zhai, Z. Ying, W. Zhou, A. Genovese, V. Piuri, and F. Scotti, "Flexible and diverse contrastive learning for steel surface defect recognition with few labeled samples," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[7] H. Zhu, J. You, Y. Zhai, Y. Xu, T. Wang, Y. Chen, J. Zhou, P. Coscia, A. Genovese, and C. L. Chen, "Ds2a-former: Battery surface defect segmentation via dual-stream spatial attention transformer network," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–17, 2025.

[8] J. Wang, H. Zou, M. Zhou, and R. Su, "An efficient deep neural network for surface defect detection in industrial edge sensing," *IEEE Trans. Ind. Inform.*, vol. 21, no. 3, pp. 2560–2569, Mar. 2025.

[9] A. Yang, X. Xu, Y. Wu, and H. Liu, "Reverse distillation for continuous anomaly detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[10] Y. Shi, B. Gao, G. Yang, H. Li, and W. L. Woo, "Industrial anomaly detection system: A multicase algorithm leveraging feature information and memory bank," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–9, 2025.

[11] X. Li, J. Jing, J. Bao, P. Lu, Y. Xie, and Y. An, "Otb-aae: Semi-supervised anomaly detection on industrial images based on adversarial autoencoder with output-turn-back structure," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[12] Y. Guo, M. Jiang, Q. Huang, Y. Cheng, and J. Gong, "Mldfr: A multilevel features restoration method based on damaged images for anomaly detection and localization," *IEEE Trans. Ind. Inform.*, vol. 20, no. 2, pp. 2477–2486, Feb. 2024.

[13] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 14 298–14 308.

[14] J. Zhu, P. Yan, J. Jiang, Y. Cui, and X. Xu, "Asymmetric teacher–student feature pyramid matching for industrial anomaly detection," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[15] Y. Zhou, X. Xu, J. Song, F. Shen, and H. T. Shen, "Msflow: Multiscale flow-based framework for unsupervised anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2437–2450, Feb. 2025.

[16] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 9659–9669.

[17] V. Zavrtanik, M. Kristan, and D. Skočaj, "DrÆm – a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8310–8319.

[18] Y. Cao, X. Xu, Z. Liu, and W. Shen, "Collaborative discrepancy optimization for reliable image anomaly localization," *IEEE Trans. Ind. Inform.*, vol. 19, no. 11, pp. 10 674–10 683, Nov. 2023.

[19] C. Chen, Q. Wu, J. Zhang, H. Xia, P. Lin, Y. Wang, M. Tian, and R. Song, "U2d2pcb: Uncertainty-aware unsupervised defect detection on pcb images using reconstructive and discriminative models," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–10, 2024.

[20] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y.-F. Wang, "Registration based few-shot anomaly detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 303–319.

[21] B. Zhu, Z. Gu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "Adformer: Generalizable few-shot anomaly detection with dual cnn-transformer architecture," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[23] T. Wang, Z. Li, Y. Xu, Y. Zhai, X. Xing, K. Guo, P. Coscia, A. Genovese, V. Piuri, and F. Scotti, "Clip-vision guided few-shot metal surface defect recognition," *IEEE Trans. Ind. Inform.*, vol. 21, no. 5, pp. 4273–4284, May 2025.

[24] Y. Liu, Q. Li, Z. Wang, J. Kato, J. Zhang, and W. Wang, "Leclip: Boosting zero-shot anomaly detection with local enhanced clip," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–11, 2025.

[25] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi, "Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2024, pp. 55–72.

[26] W. Ma, X. Zhang, Q. Yao, F. Tang, C. Wu, Y. Li, R. Yan, Z. Jiang, and S. Zhou, "Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2025, pp. 4744–4754.

[27] W. Yu, Y. Liu, X. Zhu, H. Cao, X. Sun, and X. Bai, "Turning a clip model into a scene text spotter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 6040–6054, Sep. 2024.

[28] T. Wang, Z. Li, Y. Xu, J. Chen, A. Genovese, V. Piuri, and F. Scotti, "Few-shot steel surface defect recognition via self-supervised teacher–student model with min–max instances similarity," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023.

[29] T. Liu, S. Zhou, W. Li, Y. Zhang, and J. Guan, "Semantic prototyping with clip for few-shot object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, 2025.

[30] Y. Li, E. Ivanova, and M. Bruveris, "Fade: Few-shot/zero-shot anomaly detection engine using large vision-language model," in *Proc. 35th British Mach. Vis. Conf. (BMVC)*, Nov. 2024.

[31] X. Chen, Y. Han, and J. Zhang, "A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad," *arXiv preprint arXiv:2305.17382*, 2023.

[32] J. Zhu and G. Pang, "Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 17 826–17 836.

[33] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2024.

[34] H. Yang, Z. Liu, N. Ma, X. Wang, W. Liu, H. Wang, D. Zhan, and Z. Hu, "Csrm-mim: A self-supervised pretraining method for detecting catenary support components in electrified railways," *IEEE Trans. Transp. Electrific.*, vol. 11, no. 4, pp. 10 025–10 037, 2025.

[35] J. Yan, Y. Cheng, F. Zhang, M. Li, N. Zhou, B. Jin, H. Wang, H. Yang, and W. Zhang, "Research on multimodal techniques for arc detection in railway systems with limited data," *Struct. Health Monit.*, p. 14759217251336797, 2025.

[36] S. Zhuang, H. Zhang, D. Liang, H. Liu, and Z. Gao, "Adaptive sequential bayesian iterative learning for myocardial motion estimation on cardiac image sequences," *IEEE Trans. Med. Imaging*, 2025, early Access.

[37] H. Wang, Y. Song, H. Yang, and Z. Liu, "Generalized koopman neural operator for data-driven modelling of electric railway pantograph-catenary systems," *IEEE Trans. Transp. Electrific.*, pp. 1–1, 2025.

[38] Y. Zhai, W. Pan, Y. Liang, H. Zhu, Z. Long, P. Coscia, A. Genovese, V. Piuri, and F. Scotti, "Bidirectional feature pyramid siamese anomaly

detection network with cellular anomaly generation for container marking," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–17, 2025.
[39] X. Shen, L. Li, Y. Ma, S. Xu, J. Liu, Z. Yang, and Y. Shi, "Vlcim: A vision-language cyclic interaction model for industrial defect detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–13, 2025.
[40] Z. Yang, J. Wang, X. Ye, Y. Tang, K. Chen, H. Zhao, and P. H. S. Torr, "Language-aware vision transformer for referring segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5238–5255, Jul. 2025.
[41] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 19 606–19 616.
[42] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "Anomalygpt: Detecting industrial anomalies using large vision-language models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 3, 2024, pp. 1932–1940.
[43] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9584–9592.
[44] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 13690, 2022, pp. 392–408.
[45] S. Damm, M. Laszkiewicz, J. Lederer, and A. Fischer, "Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 1319–1329.
[46] X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma, "Promptad: Learning prompts with only normal samples for few-shot anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16 848–16 858.
[47] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3606–3613.
[48] M. Baugh, J. Tan, J. P. Müller, M. Dombrowski, J. Batten, and B. Kainz, "Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks," in *Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, vol. 14220, 2023, pp. 162–172.
[49] H. M. Schlüter, J. Tan, B. Hou, and B. Kainz, "Natural synthetic anomalies for self-supervised anomaly detection and localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 13691, 2022, pp. 474–489.
[50] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.
[51] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16 795–16 804.
[52] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *Proc. Int. Symp. Ind. Electron. (ISIE)*, Jun. 2021, pp. 01–06.
[53] S. Jezek, M. Jonak, R. Burget, P. Dvorak, and M. Skotak, "Deep learning-based defect detection of metal parts: Evaluating current methods in complex conditions," in *Proc. 13th Int. Congr. Ultra Mod. Telecommun. Control Syst. Workshops (ICUMT)*. IEEE, Oct. 2021, pp. 66–71.
[54] Y. Chen, Y. Zhou, D. Lin, S. Tang, K. Tan, Y. Zhang, K. Feng, K. Hu, Y. Xu, and Y. Zhai, "Performance evaluation of anomaly detection with a novel battery spot welding anomaly dataset," in *Proc. Int. Conf. Signal Process. Neural Netw. Appl. (SPNNA)*, vol. 13652. SPIE, 2025, pp. 150–154.
[55] Y. Chen, K. Feng, Y. Xu, T. Wang, and Y. Zhai, "Unsupervised hierarchical denoising guided student-teacher framework for battery spot welding anomaly detection," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. (CIVEMSA)*, 2025, pp. 1–6.
[56] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but differnet: Semi-supervised defect detection with normalizing flows," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*.
[57] W. Luo, Y. Cao, H. Yao, X. Zhang, J. Lou, Y. Cheng, W. Shen, and W. Yu, "Exploring intrinsic normal prototypes within a single image for universal anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 9974–9983.

Yingwen Chen received the B.S.degree in Big Data Management and Application from Neusoft Institute Guangdong in 2023. He is currently working toward the M.S. degree with the School of Electronics and Information Engineering, Wuyi University.

His research interests include computer vision, pattern recognition, and image processing.

Ying Xu received the BS. and MS. degrees in automation, and control theory and control engineering from the Wuhan University of Science and Technology, Wuhan, China, in 2004 and 2008, respectively, and the Ph.D. degree in control theory and control engineering from the South China University of Science and Engineering, Guangzhou, China, in 2013.

She is currently an Associate Professor with the School of information and communication engineering, Wuyi University, Jiangmen, Guangdong, China. Her current research interests include image processing, deep learning, and biometric extraction.

Tianlei Wang received the B.S.E. degree in electrical engineering (academic training) from the Beijing University of Posts and Telecommunications, Beijing, China, in 2003, the M.S. degree in signal processing from the South China University of Technology, Guangzhou, China, in 2007, and the Ph.D. degree in safety engineering from Beijing Jiaotong University, Beijing, in 2022.

His research interests include the intelligent systems, control theory, and pattern recognition.

Yikui Zhai (Senior Member, IEEE) received his Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013.

Since October 2007, he has been working with Wuyi University, Jiangmen, China, where he is a Full Professor now. He is also Associate Dean of the School of Electronics and Information Engineering in Wuyi University, since 2021. He has been a Visiting Scholar with Department of Computer Science, the Universita degli Studi di Milano, Italy, during June 2016 to June 2017, August 2023 and January 2024. His research interests include: Image Processing, Deep Learning, Optical Character Recognition, Object Detection, UAV Change Detection, Self Supervise Learning.

Kanghong Tan received the B.S.degree in Data Science and Big Data Technology from Guangzhou College of Commerce in 2022. He is currently working toward the M.S. degree with the School of Electronics and Information Engineering, Wuyi University.

His research interests include computer vision, anomaly detection, and image processing.

Jianhong Zhou (Member, IEEE) is currently pursuing the Ph.D. degree in the School of Electronics and Information Engineering, South China University of Technology. He received the B.S. and MS. degree in Wuyi University of Communication Engineering and Information and Communication Engineering in 2020 and 2024.

His research interests include computer vision, representational contrastive learning, image processing and object detection.

Pasquale Coscia (Senior Member, IEEE) received the Ph.D. degree in Industrial and Information Engineering from the Universita degli Studi della Campania Luigi Vanvitelli, Italy, in 2019.

From 2019 to 2022, he was a post-doctoral researcher at the Universita degli Studi di Padova, Italy. He is currently an Assistant Professor at the Department of Computer Science of the Universita degli Studi di Milano, Italy. He is Co-chair of the Intelligent Measurement Systems Technical Committee (TC-22) of the IEEE Instrumentation and Measurement Society since 2023. His research activities are focused on theoretical, methodological, and applied aspects of computational intelligence for signal and image processing.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Universita degli Studi di Milan, Crema, Italy, in 2014. He has been a postdoctoral Research Fellow in computer science with the Universita degli Studi di Milan since 2014.

He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters.

His current research interests include signal and image processing, three-dimensional reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Dr. Genovese is an Associate Editor for the Journal of Ambient Intelligence and Humanized Computing (Springer).

C. L. Philip Chen (Life Fellow, IEEE) received the M.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1988, both in electrical and computer science.

He is the Chair Professor and the Dean of the College of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is the former Dean of the Faculty of Science and Technology, University of Macau, Macau, China. His research interests include cybernetics, computational intelligence, and systems.

Dr. Chen is a Fellow of the IEEE, AAAS, IAPR, CAA, and HKIE; and a member of Academia Europaea (AE) and the European Academy of Sciences and Arts (EASA). He received the IEEE Norbert Wiener Award in 2018 for his contributions in systems, cybernetics, and machine learning. He was also recognized as a Highly Cited Researcher by Clarivate Analytics from 2018 to 2021.He served as the Editor-in-Chief of IEEE Transactions on Cybernetics from 2020 to 2021, after completing his term as Editor-in-Chief of IEEE Transactions on Systems, Man, and Cybernetics: Systems from 2014 to 2019. He was President of the IEEE Systems, Man, and Cybernetics Society from 2012 to 2013. Currently, he serves as a Deputy Director of CAAI Transactions on AI, and as an Associate Editor of IEEE Transactions on AI, IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE Transactions on Fuzzy Systems, and China Sciences: Information Sciences. He has received the Macau FDCT Natural Science Award three times and the First-Rank Guangdong Province Science and Technology Advancement Award in 2019.