RWAD: Reliability-Guided Text-Free Zero-Shot Wafer Anomaly Detection

Haowen Lan*
Nanjing University
Nanjing, China

Tianrui Zhang*
Nanjing University
Nanjing, China

Ziheng Zang
Nanjing University
Nanjing, China

Silin Chen
Nanjing University
Nanjing, China

Kangjian Di
Nanjing University
Nanjing, China

Andy Liu
Nanjing University
Nanjing, China

Guohao Wang
ZetaTech Co., Ltd.
Shanghai, China

Wenzheng Zhao
ZetaTech Co., Ltd.
Shanghai, China

Ningmu Zou†
Nanjing University
Nanjing, China
nzou@nju.edu.cn

Abstract

In integrated circuit manufacturing, wafer anomaly detection is of great importance for defect analysis and yield improvement. However, wafer images in real industrial scenarios often exhibit complex textures and strong background interference, posing significant challenges to automated detection. Existing methods largely rely on manually annotated samples, category priors, or text-guided knowledge, which suffer from difficulties in data acquisition, high annotation costs, and limited cross-scenario generalization. To address these issues, we propose the first zero-shot wafer defect detection method that does not require text input, termed RWAD. Without requiring anomaly category annotations or textual prompts, the proposed method directly reconstructs normal features from wafer images to effectively perceive and identify anomalous regions. Specifically, we design a reliability-aware normal pattern extraction mechanism, which estimates patch-level normal reliability by jointly considering local neighborhood consistency and prototype distance, thereby suppressing the interference of suspicious regions in normal pattern modeling. Building upon this, we design a Normal-Pattern-Guided Dual-branch Decoder and further propose a dilated long-range remote-context decoding strategy, which enables suspicious regions to bypass locally contaminated neighborhoods, retrieve contextual information from distant high-reliability normal regions, and adaptively fuse such information with refined normal patterns to enhance the discriminability between normal and anomalous regions. Experimental results demonstrate that, even without annotations or textual priors, the proposed method can effectively adapt to complex wafer scenarios and achieve promising detection performance and generalization capability.

*Haowen Lan and Tianrui Zhang contributed equally to this work.
†Corresponding author.

ICCAD '26, San Jose, CA, USA

ACM ISBN 979-8-4007-2873-0/2026/11
https://doi.org/10.1145/3831252.3834188

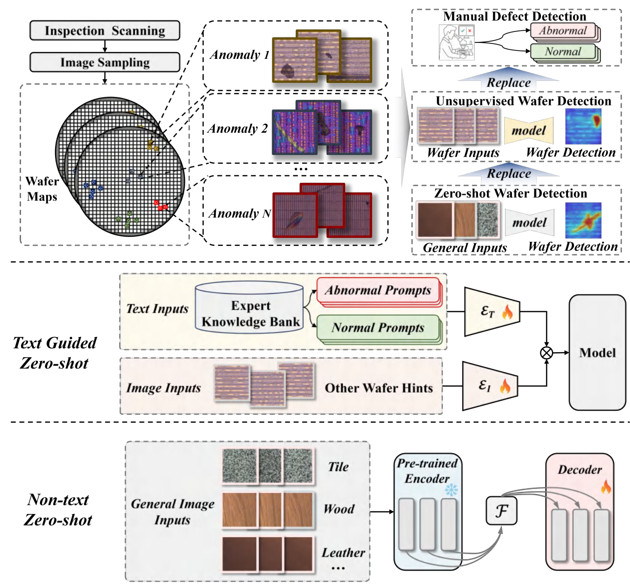

Figure 1: Evolution of wafer defect detection paradigms, illustrating the transition from manual inspection to automatic defect detection and from text-guided to text-free anomaly detection methods.

1 Introduction

Wafer manufacturing is a critical process in which high-purity silicon wafers are fabricated into semiconductor chips. This process is highly complex, and various defects may arise during manufacturing. However, even minor defects can impair chip performance and reliability and significantly reduce product yield. Therefore, reliable defect detection methods are essential for ensuring wafer quality [24]. Accordingly, extensive industrial research has focused on wafer inspection and defect detection [11, 16, 18].

Current wafer defect detection methods can generally be divided into supervised and unsupervised approaches. A large number of studies have adopted supervised learning frameworks for wafer defect detection [5, 33, 34]. These methods typically require large amounts of data and fine-grained annotations to support classification, localization, or segmentation tasks, and they often achieve strong detection performance. However, their effectiveness in standard evaluation settings comes at the cost of a heavy reliance on annotated data. In industrial applications, obtaining high-quality annotations is not only expensive but also time-consuming. Moreover, as new defect types continually emerge, existing models often struggle to transfer directly to new task scenarios.

To address this issue, industrial defect detection has increasingly shifted toward unsupervised approaches [9, 21, 22, 33]. Unsupervised detection methods require only normal samples for training and aim to automatically identify samples that deviate from normal patterns without anomaly labels or manual annotations. Since normal data usually constitute the majority and exhibit relatively stable distribution characteristics, the model can first learn the structure, distribution patterns, or feature representations of normal samples, and then regard samples with large deviations from normal patterns as anomalies. Compared with supervised methods, unsupervised anomaly detection is more suitable for real-world scenarios where anomalous samples are scarce, unknown, or difficult to annotate. Nevertheless, unsupervised learning also faces substantial challenges. Industrial data often exhibit complex and unstable structures and are easily affected by noise, anomalous samples, and redundant information. At the same time, significant inter-sample variation and diversity further complicate the task. As a result, conventional unsupervised anomaly detection strategies still struggle to fully meet the demands of complex semiconductor manufacturing scenarios.

In recent years, with the rapid development of large-scale pre-trained models, vision-language pre-training models represented by CLIP [29] have demonstrated strong cross-modal semantic alignment capabilities and have attracted widespread attention in zero-shot and few-shot recognition tasks [6, 17, 19]. By leveraging the knowledge accumulated during pre-training, such models provide a new perspective for alleviating data scarcity in industrial scenarios. However, text-guided methods still suffer from evident limitations [36, 37]. Vision-language models tend to focus more on modeling the class semantics of foreground objects rather than the abnormality or normality of the image itself. Meanwhile, for each sample, hundreds of corresponding text descriptions often need to be generated, which not only incurs high construction costs but also increases the complexity of the method. For wafer images, the differences among samples are large and the feature patterns are highly complex, making it difficult to use accurate text to comprehensively and precisely describe them. This further leads to a clear mismatch between textual semantics and defect-related visual features, limiting the applicability and generalization capability of text-guided methods in wafer defect detection tasks.

Therefore, addressing the above issues, removing the dependence on text prompts, and directly learning image representations from visual features have become important goals in zero-shot wafer defect detection. Based on this motivation, we propose RWAD, the first zero-shot wafer defect detection method that does not require text input. Compared with cross-modal methods that rely on linguistic priors, text-free zero-shot anomaly detection pays more attention to distributional differences in feature space and avoids the performance limitations caused by inaccurate textual descriptions, insufficient semantic coverage, and modality gaps. This paradigm is more suitable for handling fine-grained, weak-semantic, and hard-to-describe anomalous regions in semiconductor manufacturing images. Our method relies only on priors derived from normal samples and visual representations, allowing effective identification and localization of unknown wafer defects. The main contributions of this paper are summarized as follows:

- We propose RWAD, the first text-free zero-shot anomaly detection framework for wafer inspection. It directly models normal patterns from visual features and detects anomalies through reconstruction errors.
- We design a reliability-aware normal pattern extraction mechanism that estimates patch-level reliability using local neighborhood consistency and prototype distance. It suppresses suspicious regions during prototype aggregation to learn cleaner and more stable normal representations.
- We design a Normal-Pattern-Guided Dual-branch Decoder with a dilated long-range remote-context decoding strategy. It avoids potentially contaminated local neighborhoods and retrieves contextual information from distant high-reliability normal regions, improving feature reconstruction and anomaly localization.
- Experiments demonstrate that RWAD achieves strong detection performance and generalization without anomaly annotations or textual priors.

2 Related Work

2.1 Wafer Surface Defect Detection

Wafers serve as the fundamental substrate for integrated circuit manufacturing, and their surface quality directly affects chip performance, reliability, and final yield. However, wafer surface images usually exhibit highly repetitive structural textures and complex background interference, while defect regions are often subtle, sparse, and morphologically diverse, making defect detection and localization highly challenging.

Early wafer defect detection methods mainly relied on traditional machine learning or supervised deep learning frameworks. For example, early approaches performed defect detection by combining handcrafted features with conventional classifiers. Gómez-Sirvent et al. employed a feature-selection-based support vector machine framework for semiconductor wafer defect classification [12]. With the development of deep learning, convolutional neural network based models have been widely adopted for wafer defect classification and segmentation. For instance, Cheon et al. proposed a CNN framework for wafer surface defect classification and unknown defect category detection [7], Saqlain et al. designed a deep convolutional network for wafer defect identification under class-imbalanced scenarios [31], and de la Rosa et al. combined a lightweight SqueezeNet architecture for semiconductor wafer defect detection and classification [8]. Although these methods have achieved promising results, they generally depend on sufficient training samples and accurate pixel-level annotations, and thus often show limited effectiveness when data are scarce or new defect types emerge.

In real semiconductor manufacturing environments, anomalous samples are scarce and costly to collect, while defect patterns may vary significantly across different product layouts and process conditions, which further limits the applicability of supervised methods. Therefore, many unsupervised anomaly detection methods for industrial inspection have recently emerged. For example, FD-UAD [3] enhances defect perception in complex industrial scenarios through a multimodal imaging-based unsupervised anomaly detection framework, while GLASS [4] and RD++ [32] improve industrial anomaly detection and localization from the perspectives of anomaly synthesis and reverse distillation, respectively. Nevertheless, existing unsupervised wafer anomaly detection methods still face difficulties in detecting subtle defects under complex backgrounds and in modeling normal patterns when appearance variations across categories are large. Furthermore, under the zero-shot setting, the distribution of normal features across different wafer categories may shift significantly, which further increases the difficulty of stable normal pattern modeling. Therefore, there is an urgent need to develop an automated framework that can effectively identify and localize wafer defects under limited-data or even target-domain annotation-free settings.

2.2 Few-shot/Zero-shot Anomaly Detection Methods

With the increasing scarcity of anomalous samples and the high cost of annotation in industrial scenarios, few-shot and zero-shot anomaly detection have attracted growing attention. These methods aim to identify and localize anomalous samples with very limited or even no anomaly annotations in the target domain, thereby alleviating the dependence of traditional supervised methods on large-scale labeled data and providing a more flexible and efficient solution to the data scarcity problem in real-world industrial environments.

Existing few-shot and zero-shot anomaly detection methods can generally be divided into two categories. One category is based on pretrained vision-language models, which align image features with normal/abnormal semantics through text prompts, fine-grained descriptions, or anomaly-aware adaptation mechanisms. For example, AdaCLIP, FiLo, AA-CLIP, and PA-CLIP improve zero-shot anomaly detection and localization from the perspectives of hybrid prompt learning, fine-grained anomaly description, anomaly-aware adaptation, and prompt adaptation, respectively [2, 14, 25, 27], while Anomaly-OV further enhances zero-shot anomaly detection and reasoning with multimodal large models [35]. The other category focuses more on leveraging a small number of normal samples for contextual modeling or directly characterizing normal patterns, and treats regions deviating from the normal distribution as potential anomalies. For instance, InCTRL achieves cross-domain anomaly detection through few-shot normal sample prompts and in-context residual learning [38], One-to-Normal enhances adaptability in few-shot scenarios from the perspective of anomaly personalization [20], and INP-Former further improves anomaly detection performance with a stronger normal prototype modeling mechanism [23].

Although these methods have achieved promising results on general industrial anomaly detection tasks, they still exhibit clear limitations in wafer defect detection scenarios. Text-based methods usually require constructing a large number of corresponding textual descriptions for each sample, while the large variations among wafer samples make it difficult to describe them accurately with text. Methods based on normal pattern modeling, on the other hand, usually rely on modeling the normal distribution, yet wafer images are difficult to represent in a stable and accurate manner. Meanwhile, defect signals can easily interfere with the normal distribution, which further limits the recognition accuracy and generalization capability of such methods in wafer defect detection tasks. Therefore, how to stably characterize the normal patterns of wafer images without relying on complex text prompts, while improving robust recognition and localization of subtle defects, remains a key challenge in zero-shot wafer anomaly detection.

3 Method

3.1 RWAD Framework

In this section, we propose the RWAD framework for zero-shot wafer defect detection without textual prompts. Unlike methods that rely on category descriptions or manually designed prompts, RWAD learns normal patterns directly from image features and suppresses anomalous noise through reliability-guided normal pattern modeling, which leads to more stable defect response maps.

The overall framework consists of three components:

- a pretrained visual encoder for extracting multi-level patch token features;
- a Reliability-Guided Normal Pattern Extractor, which learns a set of prototypes to identify suspicious tokens and reduce their influence during normal pattern modeling in wafer images;
- a normal-pattern-guided decoder and a reliability-guided remote context decoder for reconstructing normal structures and generating anomaly maps.

Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we first feed it into a pretrained visual encoder Q to extract multi-level feature representations and introduce a set of learnable prototypes:

$$\mathcal{F}_Q = \left\{\mathbf{F}_Q^{(l)} \in \mathbb{R}^{N\times C}\right\}_{l=1}^{L}, \quad N = \frac{HW}{k^2}, \qquad \mathcal{P} = \{\mathbf{p}_m \in \mathbb{R}^C\}_{m=1}^{M}. \tag{1}$$

H and W denote the input height and width, respectively; L is the number of feature levels, C is the channel dimension, k is the downsampling factor, and M is the number of prototype tokens. Based on the encoded features and the learnable prototypes, we estimate a reliability score for each token according to its local consistency and prototype distance. Let $\mathbf{r} = \{r_i\}_{i=1}^N$ denote the token-wise reliability set. Then, the refined normal patterns (RNPs) are obtained by

$$\mathcal{P}^r = \mathrm{Agg}(\mathcal{P}, \mathcal{F}_Q; \mathbf{r}), \tag{2}$$

where $\mathcal{P}^r$ denotes the refined normal patterns and $\mathrm{Agg}(\cdot)$ is the reliability-guided prototype aggregation operation.

After that, the refined normal patterns $\mathcal{P}^r$ are fed into a normal-pattern-guided decoder D_p, while the encoded features $\mathcal{F}_Q$ together with the token-wise reliability $\mathbf{r}$ are fed into a reliability-guided

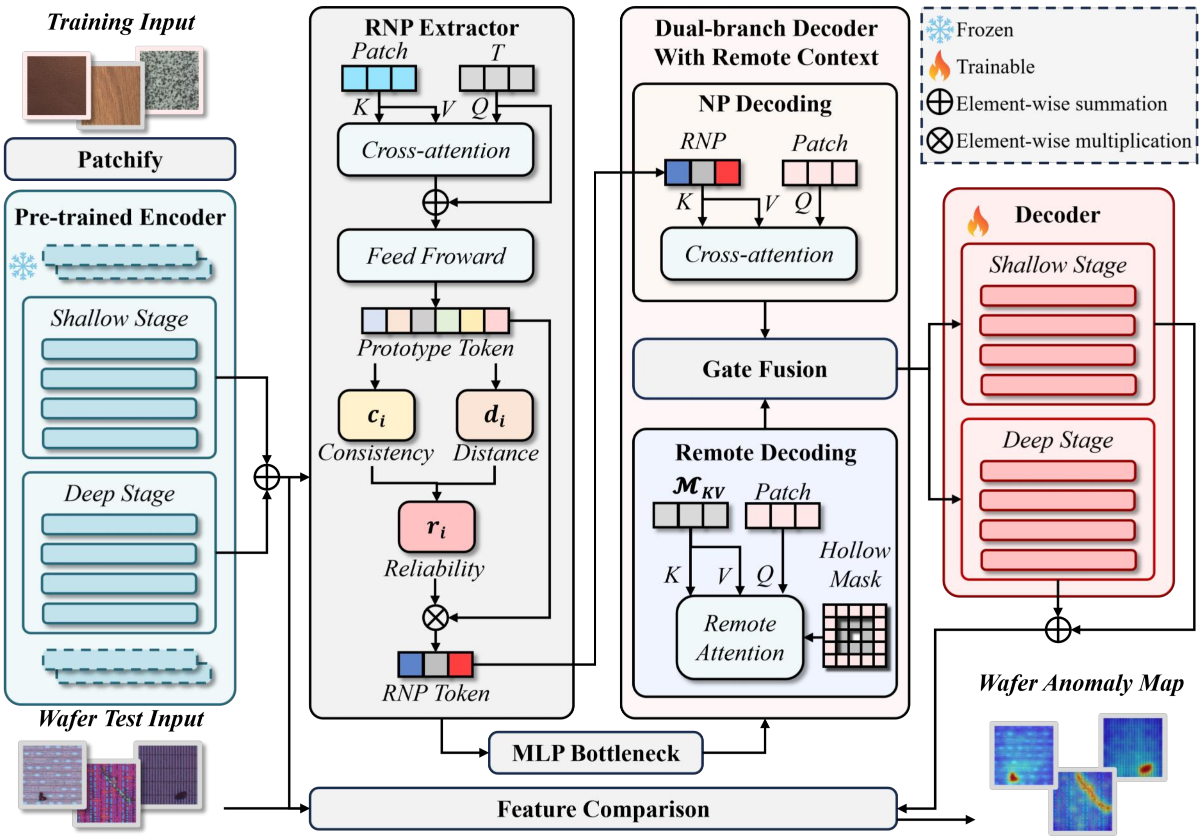

Figure 2: Overall framework of the proposed RWAD.

remote-context decoder D_c:

$$\mathcal{F}_D^p = D_p(\mathcal{P}^r), \qquad \mathcal{F}_D^c = D_c(\mathcal{F}_Q, \mathbf{r}). \tag{3}$$

Finally, the outputs of the two branches are fused as

$$\mathcal{F}_D = \mathcal{F}_D^p + (1 - \mathbf{r}) \odot \mathcal{F}_D^c \tag{4}$$

where $\odot$ denotes element-wise multiplication. The anomaly map is finally obtained by measuring the discrepancy between the encoded features and the reconstructed features. During inference, we directly load the model parameters trained on the source domain and transfer them to the target wafer dataset without requiring any target-domain textual descriptions or additional annotations.

3.2 Visual Feature Encoding

Here, we adopt a pretrained DINOv2 ViT encoder [26] and keep it frozen to preserve the generic feature representation capability learned from large-scale natural image data, while reducing the optimization difficulty caused by additional trainable parameters. As the visual feature extraction backbone, the frozen encoder can stably produce hierarchical patch token representations, providing unified feature inputs for the subsequent reliability-guided normal pattern modeling and dual-branch decoder reconstruction.

Given an input wafer image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we feed it into a pretrained DINOv2 ViT encoder Q to extract visual representations with hierarchical semantic information. Instead of using only the last-layer feature, we select the outputs of several intermediate Transformer blocks as feature sources, so as to preserve both shallow texture details and deep semantic context. The output of the l-th selected layer is $\mathbf{F}_Q^{(l)} = Q^{(l)}(\mathbf{I}) \in \mathbb{R}^{N \times C}$, where N denotes the number of patch tokens and C denotes the channel dimension. Since the [CLS] token does not directly participate in pixel-level anomaly localization, we remove it in the subsequent processing and retain only the patch token representations.

To fuse visual information from different levels, we average the token features of the selected layers to obtain the final encoded feature representation:

$$\mathcal{F}_Q = \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \mathbf{F}_Q^{(l)}, \tag{5}$$

where $\mathcal{S}$ denotes the set of selected intermediate layers. This strategy effectively integrates shallow local textures with deep global semantic information, thereby enhancing the model's representation capability for defects of different scales.

3.3 Reliability-Guided Normal Pattern Extractor

Existing anomaly detection methods based on normal reference modeling [13, 28, 30] can generally be divided into two categories.

The first stores local normal representations from training samples and performs token-level matching with the test image. The second dynamically derives global normal cues from the test image itself. However, such dynamic extraction strategies usually operate on the entire image directly, making the learned normal references vulnerable to contamination by anomalous regions and noisy patterns, which may in turn degrade detection accuracy. To address this issue, our approach jointly incorporates local consistency and prototype distance to estimate wafer-specific normal reliability, thereby suppressing the influence of suspicious features during normal pattern modeling.

To capture stable normal patterns in wafer images, RWAD maintains M learnable prototype tokens, denoted by $\mathcal{P} = \{\mathbf{p}_m \in \mathbb{R}^C\}_{m=1}^M$. To adapt these prototypes to the current input, we employ prototype-to-image cross-attention, with prototypes as queries and image tokens as keys and values, to produce coarse pattern embeddings $\widetilde{\mathcal{P}} = \mathrm{Agg}(\mathcal{P}, \mathcal{F}_Q)$.

Since anomalous or noisy regions may still interfere with normal pattern modeling, we further estimate the reliability of each token from two complementary perspectives: local neighborhood consistency and prototype distance. For the i-th token, these two quantities are computed as

$$c_i = \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \cos(\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j), \qquad d_i = 1 - \max_m \cos(\hat{\mathbf{f}}_i, \hat{\tilde{\mathbf{p}}}_m), \tag{6}$$

where $\mathcal{N}(i)$ denotes the neighborhood centered at token i, $\hat{\mathbf{f}}_i$ is the normalized token feature, and $\hat{\tilde{\mathbf{p}}}_m$ denotes the normalized coarse pattern embedding. Intuitively, normal tokens are expected to be more consistent with their neighbors and more compatible with the learned pattern set. These two cues are then combined to form a suspiciousness score:

$$s_i = \alpha(1 - c_i) + \beta d_i, \tag{7}$$

where α and β control the relative contributions of local consistency and prototype distance. Based on this score, the reliability weight is defined as

$$r_i = 1 - \sigma(\gamma s_i + \delta). \tag{8}$$

In this way, more suspicious tokens are assigned lower reliability values. Using the token-wise reliability weights $\mathbf{r} = \{r_i\}_{i=1}^N$, we perform reliability-guided aggregation to obtain the refined normal patterns (RNPs):

$$\mathcal{P}^r = \mathrm{Agg}(\mathcal{P}, \mathcal{F}_Q; \mathbf{r}). \tag{9}$$

Compared with the coarse pattern embeddings, $\mathcal{P}^r$ is less affected by suspicious regions and therefore provides a cleaner and more stable representation of the normal structures in the input wafer image.

3.4 Normal-Pattern-Guided Dual-Branch Decoder with Remote Context

After obtaining the refined normal patterns $\mathcal{P}^r$, we further construct a dual-branch decoding mechanism to reconstruct the encoded token features into representations that are closer to normal patterns. Specifically, the decoder consists of a normal-pattern-guided branch and a reliability-guided remote-context branch. The former uses the RNPs as structural priors to guide token reconstruction toward the normal distribution, while the latter introduces long-range contextual cues to compensate for insufficient local information and suppresses the influence of unreliable regions through reliability weighting. Through the collaboration of the two branches, the model can more stably recover normal structures under complex wafer textures and enlarge the discrepancy between anomalous regions and normal patterns.

Let $\mathcal{F}_Q$ denote the fused feature representation from the encoder. Before decoding, a lightweight bottleneck produces $\mathbf{H}^0 = \mathrm{MLP}(\mathcal{F}_Q) \in \mathbb{R}^{N\times C}$. Starting from $\mathbf{H}^0$, the normal-pattern-guided branch interacts with the RNPs $\mathcal{P}^r \in \mathbb{R}^{M\times C}$ to generate normal-pattern-constrained reconstruction features. At stage t, the token features $\mathbf{H}^t$ and RNPs are normalized and linearly projected into queries $\mathbf{Q}_p$, keys $\mathbf{K}_p$, and values $\mathbf{V}_p$. Instead of standard softmax attention, their similarity is reweighted by a learnable head-wise scale λ_p and passed through a non-negative activation, yielding $\mathbf{A}_p = \mathrm{ReLU}((\mathbf{Q}_p\mathbf{K}_p^\top) \odot \lambda_p)$. The value embeddings are then aggregated and refined by a feed-forward network:

$$\mathbf{H}_p^{t+1} = \mathbf{A}_p\mathbf{V}_p + \mathrm{FFN}\Big(\mathrm{Norm}\big(\mathbf{A}_p\mathbf{V}_p\big)\Big). \tag{10}$$

In this manner, token representations are explicitly constrained by the RNPs, encouraging the reconstructed features to approach normal patterns.

Although the normal-pattern-guided branch is effective for normal pattern modeling, relying solely on the RNPs may still be insufficient in wafer images with repeated textures and strong local similarity. Many anomalous regions remain locally similar to normal ones, making them difficult to distinguish based only on nearby cues. To alleviate this issue, we further introduce a reliability-guided remote-context branch, which retrieves complementary information from long-range spatial relations.

Given the current token features $\mathbf{H}^t \in \mathbb{R}^{N\times C}$ and remote memory $\mathbf{M}^t \in \mathbb{R}^{N\times C}$, we normalize and linearly project them into $\mathbf{Q}_r$, $\mathbf{K}_r$, and $\mathbf{V}_r$. A hollow mask Ω blocks neighboring positions while preserving long-range context, and the reliability weights $\mathbf{r} \in [0,1]^N$ suppress unreliable responses. The resulting remote attention is

$$\mathbf{A}_r = \mathrm{Softmax}\left(\frac{\mathrm{Norm}(\mathbf{Q}_r)\mathrm{Norm}(\mathbf{K}_r)^\top}{\sqrt{d}} + \log(\mathbf{r}) + \Omega\right). \tag{11}$$

Its value aggregation and residual enhancement are jointly written as

$$\mathbf{H}_r^{t+1} = \mathbf{A}_r\mathbf{V}_r + \mathrm{FFN}(\mathrm{Norm}(\mathbf{A}_r\mathbf{V}_r)). \tag{12}$$

Rather than simply enlarging the receptive field, this branch performs selective global aggregation under reliability constraints, enabling the model to seek more stable support from distant normal structures when local regions are ambiguous or unreliable. To fully exploit the complementarity of the two branches, we finally adopt a reliability-gated fusion strategy:

$$\mathbf{H}^{t+1} = \mathbf{H}_p^{t+1} + (1 - \mathbf{r}) \odot \mathbf{H}_r^{t+1}, \tag{13}$$

where $\odot$ denotes element-wise multiplication. This design makes the model rely more on normal-pattern-guided reconstruction for reliable tokens, while assigning greater importance to the remote-context branch for unreliable ones. Through such a dynamic balance between local recovery and global compensation, the decoder can better preserve normal regions while reducing the risk of over-restoring anomalous regions under locally similar textures.

Overall, the decoding process in RWAD can be viewed as a collaborative reconstruction mechanism that integrates RNP-based structural priors, remote-context compensation, and reliability-gated fusion. This design allows the model to recover normal structures more accurately while preserving more discriminative reconstruction residuals in anomalous regions, thereby providing stronger signals for subsequent anomaly map generation.

3.5 Loss Function

To optimize RWAD, we employ a joint objective consisting of a reconstruction loss and a pattern gathering loss:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{proto}, \tag{14}$$

where $\lambda = 0.2$ in all experiments. The reconstruction loss supervises the consistency between encoder features and decoder features at multiple levels:

$$\mathcal{L}_{rec} = \frac{1}{L} \sum_{l=1}^{L} \left(1 - \cos(\mathbf{E}_l, \mathbf{D}_l)\right). \tag{15}$$

To encourage the model to focus on hard-to-reconstruct tokens, we adopt an adaptive reweighting strategy during optimization. In addition, the pattern gathering loss measures the minimum cosine distance between each token feature and the learned refined normal patterns (RNPs):

$$\mathcal{L}_{proto} = \frac{1}{N} \sum_{i=1}^{N} \min_{k} \left(1 - \cos(\mathbf{f}_i, \mathbf{p}_k)\right). \tag{16}$$

This objective encourages the RNPs to capture stable normal patterns while improving the robustness of reconstruction under complex wafer textures.

4 Experiments

4.1 Experimental Settings

We evaluate RWAD from both image-level and pixel-level perspectives. For image-level anomaly classification, we adopt AUROC, AP, and F1-max to measure the ability to distinguish normal and anomalous images. For pixel-level defect localization, we also adopt AUROC, AP, and F1-max to evaluate the accuracy of localized anomaly regions.

We employ a pretrained DINOv2-reg ViT-B/14 [26] as the image encoder, with an input resolution of 392×392. Multi-level features extracted from intermediate encoder blocks are used to construct multi-scale representations. The number of learnable prototypes is set to 6. The decoder consists of a normal-pattern-guided decoder and a remote-context decoder. All experiments are conducted on an NVIDIA RTX A6000 GPU. During training, only the learnable prototypes, bottleneck projection layers, and decoder parameters are optimized, while the image encoder remains frozen. We use the AdamW optimizer with an initial learning rate of 1×10^{-3}, a weight decay of 1×10^{-4}, and a batch size of 16. The model is trained for 200 epochs with a warm cosine learning rate scheduler.

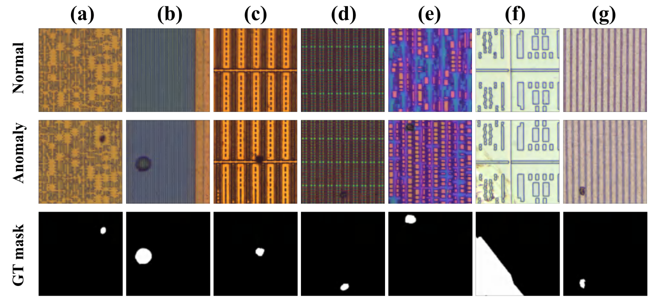

Figure 3: Normal, anomaly, and ground-truth mask examples from PWmicro-AD.

We compare these anomaly detection methods under the zero-shot setting, including CLIP [29], WinCLIP [17], PromptAD [19], DRA [10] and AnomalyGPT [15]. Among them, OriginalCLIP, WinCLIP, and PromptAD are CLIP-based methods, for which we use ViT-B/16+ pretrained on LAION-400M as the backbone.

4.2 Our Dataset

To evaluate the effectiveness of our method in realistic cross-domain industrial scenarios, we establish a zero-shot detection setting with PWmicro-AD as the target-domain benchmark. This dataset is designed for patterned wafer inspection and contains 15 categories of patterned wafer images, with a total of 7,127 high-resolution RGB samples, including 4,918 normal samples and 2,209 anomalous samples. The anomalous images cover various fine-grained defect types, such as particles and structural deformations, and all of them are equipped with pixel-level ground-truth masks, enabling evaluation at both the image level for anomaly classification and the pixel level for defect localization. Our dataset is illustrated in Fig. 3, where (a)–(g) show different types of samples in the dataset. The dataset will be made publicly available on the project website.

As the source-domain training data, we adopt the widely used MVTec AD dataset [1], which contains both texture and object categories of industrial defects and can provide general anomaly representation capability for the model. In our experiments, RWAD is trained only on MVTec and then directly deployed on PWmicro-AD for testing. During this process, no anomaly annotations from the target domain are used, no target-domain fine-tuning is performed, and no text prompts are introduced, thereby realizing a strict zero-shot wafer anomaly detection setting.

4.3 Results Analysis

From the overall results reported in Table 1, it can be observed that RWAD achieves excellent performance in both image-level detection and pixel-level localization, yielding the best overall detection results. Specifically, RWAD attains 0.938, 0.974, and 0.922 on I-AUROC, I-AP, and I-F1-max, respectively. Compared with AnomalyGPT, which is the closest competitor in image-level performance,

Table 1: Comparison of RWAD with other anomaly detection methods under the zero-shot setting.

Method	Image-level			Pixel-level		
	AUROC	AP	F1-max	AUROC	AP	F1-max
OriginalCLIP [29]	0.738	0.877	-	0.199	0.012	-
WinCLIP [17]	0.789	0.898	0.859	0.286	0.014	0.041
DRA [10]	0.619	0.794	0.824	0.637	0.068	0.116
PromptAD [19]	0.824	0.920	0.855	0.917	0.547	0.555
AnomalyGPT [15]	0.890	0.946	-	0.913	0.476	-
RWAD	**0.938**	**0.974**	**0.922**	**0.938**	**0.571**	**0.572**

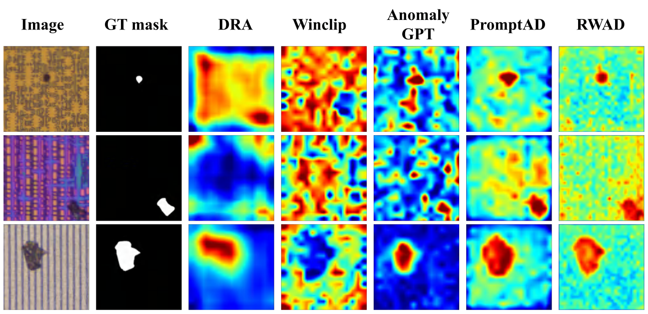

Figure 4: Visualization of method comparison under the zero-shot setting.

RWAD improves I-AUROC and I-AP by 0.048 and 0.028, respectively, further validating its advantage in image-level anomaly recognition. Meanwhile, RWAD also performs remarkably well in pixel-level localization, achieving 0.938, 0.571, and 0.572 on P-AUROC, P-AP, and P-F1-max, respectively. Compared with PromptAD, RWAD further improves P-AUROC, P-AP, and P-F1-max by 0.021, 0.024, and 0.017, respectively, indicating that the proposed method not only identifies anomalous images more accurately but also delineates defective regions more precisely. These results fully demonstrate that RWAD maintains excellent image-level detection performance while further enhancing fine-grained defect localization, thereby achieving stronger overall performance in zero-shot wafer anomaly detection. The qualitative detection results are shown in Fig. 4.

RWAD achieves these improvements mainly because it can construct cleaner normal patterns. On the one hand, the reliability-guided normal pattern extraction mechanism suppresses the interference of potentially anomalous regions during prototype learning, making the constructed normal patterns cleaner and more stable. On the other hand, the long-range remote-context decoding strategy is able to retrieve effective information from distant high-reliability normal regions even when local neighborhoods are contaminated, thereby improving reconstruction quality and enhancing the separability between normal and anomalous regions. Therefore, the leading performance of RWAD on pixel-level metrics validates the effectiveness of the proposed method for fine-grained anomaly localization.

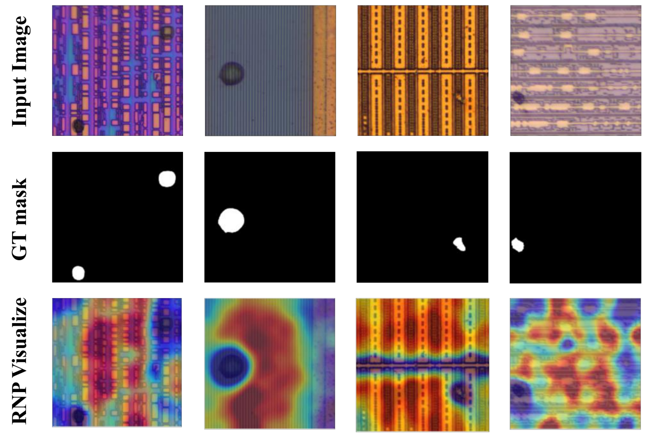

Figure 5: Visualization of RNP attention extraction.

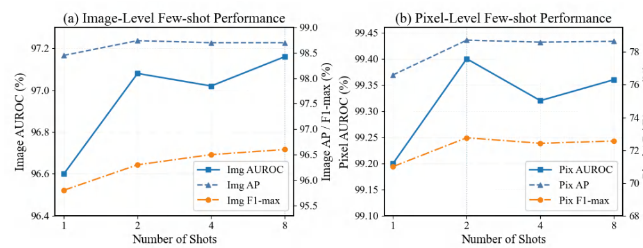

Figure 6: k-shot experimental results of RWAD (AUROC/AP/F1-max).

The visualization results in Fig. 5 further demonstrate the effectiveness of the proposed reliability-guided normal pattern extraction. It can be observed that the attention of the extracted RNPs is mainly concentrated on normal structural regions, while anomalous patches are largely suppressed. This indicates that the proposed extractor can effectively reduce the interference of suspicious regions during prototype aggregation and learn cleaner normal representations, which is beneficial for subsequent reconstruction and anomaly localization.

Furthermore, to further validate the effectiveness and stability of RWAD, we conduct experiments under different k-shot settings, as shown in Fig. 6. The results show that RWAD still maintains strong detection performance even under extremely low-shot conditions. Under the 1-shot setting, the model already achieves high image-level and pixel-level performance. When the shot number increases to 2, the performance is further improved, especially on the pixel-level metrics, which increase from 0.992/0.766/0.711 to 0.994/0.787/0.728. Meanwhile, the performance gap among the 2-shot, 4-shot, and 8-shot settings is already small, indicating that RWAD can learn stable and representative normal patterns with only a few normal samples.

Combined with the comparative results discussed above, RWAD not only achieves strong performance in both image-level detection and pixel-level localization, but also demonstrates good robustness under low-shot settings. In particular, under the 2-shot

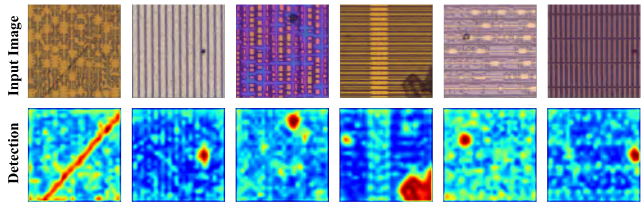

Figure 7: Detection results under the 2-shot setting.

Table 2: Ablation study of RWAD (AUROC/AP/F1-max).

RNP Extractor	Remote Context Decoder	Image-level	Pixel-level
×	×	0.876/0.940/0.892	0.906/0.508/0.526
✓	×	0.931/0.970/0.919	0.906/0.554/0.562
✓	✓	**0.938/0.974/0.922**	**0.938/0.571/0.572**

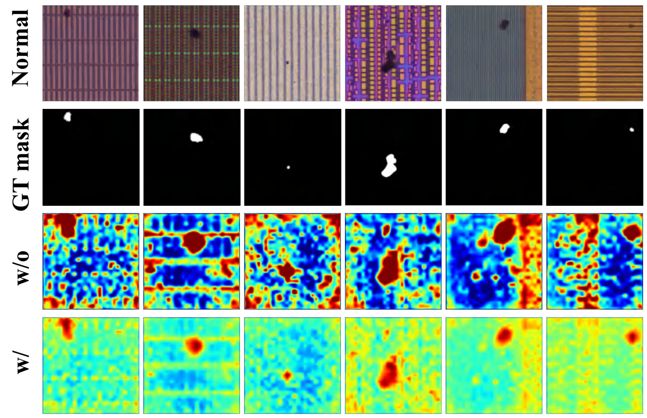

Figure 8: Detection results in the 0-shot ablation study. The third and fourth rows present the results without the RNP Extractor and Remote Context Decoder and those of the full RWAD.

setting, the model already achieves nearly optimal detection and localization performance with low data cost, which further validates the effectiveness of the reliability-guided normal pattern modeling and long-range remote-context decoding strategy in low-shot scenarios. The qualitative detection results are shown in Fig. 7.

4.4 Ablation Study

To verify the effectiveness of the proposed components, we conduct ablation studies on the RNP Extractor and the Remote Context Decoder. When both modules are removed, the model degenerates into a basic framework that employs standard cross-attention to extract coarse normal patterns and uses only the normal-pattern-guided decoder branch for feature reconstruction. All model variants are evaluated under the same zero-shot setting and evaluation protocol to ensure a fair comparison.

The ablation results are shown in Table 2. When only the RNP Extractor is introduced, both image-level and pixel-level performance improve over the baseline. The image-level AUROC, AP, and F1-max increase from 0.876/0.940/0.892 to 0.931/0.970/0.919. Although the pixel-level AUROC remains at 0.906, the pixel-level AP and F1-max increase from 0.508 and 0.526 to 0.554 and 0.562, respectively. These results indicate that the RNP Extractor can reduce the interference of suspicious regions during normal pattern aggregation and obtain cleaner and more reliable normal representations. The resulting normal patterns provide more stable guidance for subsequent feature reconstruction, improving image-level anomaly recognition and enhancing the discrimination of anomalous pixels.

After further introducing the Remote Context Decoder, the model achieves consistent improvements and obtains the best overall results. The image-level AUROC, AP, and F1-max reach 0.938, 0.974, and 0.922, respectively, while the pixel-level AUROC, AP, and F1-max reach 0.938, 0.571, and 0.572, respectively. Compared with the model using only the RNP Extractor, the complete model achieves more evident improvements in pixel-level performance. This is because the Remote Context Decoder can bypass potentially contaminated local neighborhoods and retrieve complementary information from distant regions with high normal reliability. When local structures are anomalous or provide insufficient information, such long-range normal context offers a more stable reference for feature reconstruction, reducing the influence of local contamination and further improving anomaly localization. The qualitative visualization of the ablation study is shown in Fig. 8.

Compared with the baseline without the two proposed modules, the complete model produces clearer and more concentrated responses in anomalous regions while reducing noisy responses in normal regions. The quantitative and qualitative results jointly demonstrate that the RNP Extractor and the Remote Context Decoder play different yet complementary roles in RWAD. The former constructs cleaner and more reliable normal patterns, whereas the latter further improves reconstruction quality by exploiting long-range normal context. Their combination enables the model to maintain strong image-level anomaly detection performance while achieving more accurate pixel-level defect localization.

5 Conclusion

This paper proposes RWAD, the first text-free zero-shot anomaly detection framework for industrial wafer inspection. RWAD directly models normal patterns from visual features without anomaly annotations, category priors, or textual prompts. It combines a reliability-guided normal pattern extractor with a dual-branch decoder to suppress suspicious regions and improve normal feature reconstruction. Experiments on PWmicro-AD demonstrate strong performance in both image-level anomaly detection and pixel-level defect localization, providing an effective solution for wafer inspection with scarce annotations.

References

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

[2] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. 2024. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European conference on computer vision*. Springer, 55–72.

[3] Yang Chang, Yuxuan Lin, Boyang Wang, Qing Zhao, Yan Wang, and Wenqiang Zhang. 2024. FD-UAD: Unsupervised Anomaly Detection Platform Based on Defect Autonomous Imaging and Enhancement.. In *IJCAI*. 8619–8622.

[4] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. 2024. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In *European Conference on Computer Vision*. Springer, 37–54.

[5] Shouhong Chen, Zhentao Huang, Tao Wang, Xingna Hou, and Jun Ma. 2025. Wafer map defect recognition based on multi-scale feature fusion and attention spatial pyramid pooling. *Journal of Intelligent Manufacturing* 36, 1 (2025), 271–284.

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

[7] Sejune Cheon, Hankang Lee, Chang Ouk Kim, and Seok Hyung Lee. 2019. Convolutional neural network for wafer surface defect classification and the detection of unknown defect class. *IEEE Transactions on Semiconductor Manufacturing* 32, 2 (2019), 163–170.

[8] Francisco López de la Rosa, José L Gómez-Sirvent, Rafael Morales, Roberto Sánchez-Reolid, and Antonio Fernández-Caballero. 2023. Defect detection and classification on semiconductor wafers using two-stage geometric transformation-based data augmentation and SqueezeNet lightweight convolutional neural network. *Computers & Industrial Engineering* 183 (2023), 109549.

[9] Hanqiu Deng and Xingyu Li. 2022. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9737–9746.

[10] Choubo Ding, Guansong Pang, and Chunhua Shen. 2022. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7388–7398.

[11] Hao Geng, Qi Sun, Tinghuan Chen, Qi Xu, Tsung-Yi Ho, and Bei Yu. 2023. Mixed-type wafer failure pattern recognition. In *Proceedings of the 28th Asia and South Pacific Design Automation Conference*. 727–732.

[12] José L Gómez-Sirvent, Francisco López de la Rosa, Roberto Sánchez-Reolid, Antonio Fernández-Caballero, and Rafael Morales. 2022. Optimal feature selection for defect classification in semiconductor wafers. *IEEE transactions on semiconductor manufacturing* 35, 2 (2022), 324–331.

[13] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1705–1714.

[14] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. 2024. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2041–2049.

[15] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2024. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 1932–1940.

[16] Xingna Hou, Mulan Yi, Shouhong Chen, Meiqi Liu, and Ziren Zhu. 2024. Recognition and classification of mixed defect pattern wafer map based on multi-path DCNN. *IEEE Transactions on Semiconductor Manufacturing* 37, 3 (2024), 316–328.

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

[18] Qian Jin, Yuqi Jiang, Xudong Lu, Yumeng Liu, Yining Chen, Dawei Gao, Qi Sun, and Cheng Zhuo. 2024. Sem-clip: Precise few-shot learning for nanoscale defect detection in scanning electron microscope image. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*. 1–8.

[19] Yiting Li, Adam Goodge, Fayao Liu, and Chuan-Sheng Foo. 2024. Promptad: Zero-shot anomaly detection using text prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1093–1102.

[20] Yiyue Li, Shaoting Zhang, Kang Li, and Qicheng Lao. 2024. One-to-normal: Anomaly personalization for few-shot anomaly detection. *Advances in Neural Information Processing Systems* 37 (2024), 78371–78393.

[21] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. 2023. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20402–20411.

[22] Hu Lu, Jiwei Shen, Botong Zhao, Pengjie Lou, Wenzhan Zhou, Kan Zhou, Xintong Zhao, Shujing Lyu, and Yue Lu. 2023. A masked autoencoder-based approach for defect classification in semiconductor manufacturing. In *2023 International Workshop on Advanced Patterning Solutions (IWAPS)*. IEEE, 1–4.

[23] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. 2025. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 9974–9983.

[24] Jianhong Ma, Tao Zhang, Cong Yang, Yangjie Cao, Lipeng Xie, Hui Tian, and Xuexiang Li. 2023. Review of wafer surface defect detection methods. *Electronics* 12, 8 (2023), 1787.

[25] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. 2025. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 4744–4754.

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[27] Yurui Pan, Lidong Wang, Yuchao Chen, Wenbing Zhu, Bo Peng, and Mingmin Chi. 2025. Pa-clip: Enhancing zero-shot anomaly detection through pseudo-anomaly awareness. *arXiv preprint arXiv:2503.01292* (2025).

[28] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14372–14381.

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

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

[31] Muhammad Saqlain, Qasim Abbas, and Jong Yun Lee. 2020. A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing* 33, 3 (2020), 436–444.

[32] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. 2023. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24511–24520.

[33] Qijing Wang and Martin DF Wong. 2022. Waferhsl: Wafer failure pattern classification with efficient human-like staged learning. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*. 1–8.

[34] Rui Wang, Songhao Wang, and Ben Niu. 2025. Shape prior guided defect pattern classification and segmentation in wafer bin maps. *Journal of Intelligent Manufacturing* 36, 1 (2025), 319–330.

[35] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. 2025. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 20370–20382.

[36] Tianrui Zhang, Liang Gao, Xinyu Li, and Yiping Gao. 2025. Dzad: Diffusion-based zero-shot anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10131–10138.

[37] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. 2023. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961* (2023).

[38] Jiawen Zhu and Guansong Pang. 2024. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17826–17836.