Adaptive Feature Refinement with Information Preservation for Multiclass Unsupervised Anomaly Detection

Junho Lee[1], Jincheol Yang[1], Geonwoo Kim[1], U-Yeong Kim[1], Yunseok Song[2], Hyun-Boo Lee[2], Jimin Roh[2], Heechul Lim[3], Suk-Ju Kang[1*]

[1]Sogang University, Seoul, Korea.
[2]LG Electronics, Pyeongtaek, Korea.
[3]Tech University of Korea, Siheung, Korea.

*Corresponding author(s). E-mail(s): sjkang@sogang.ac.kr;
Contributing authors: torong15@sogang.ac.kr; yjc3232@sogang.ac.kr; kimgeonwoo1998@sogang.ac.kr; drew0523@sogang.ac.kr; yunseok.song@lge.com; hyunboo.lee@lge.com; jimin.roh@lge.com; hc.lim@tukorea.ac.kr;

Abstract

Recent advancements in anomaly detection have shown significant potential across various industrial domains. However, a wide range of unpredictable defect types emerge in the real world, and anomaly images are often challenging to obtain, making traditional methods less suitable. To address this challenge, recent studies have focused on the multiclass unsupervised anomaly detection task. Nonetheless, these approaches face significant challenges owing to the difficulty in robustly handling diverse classes and defect types. We propose Adaptive Feature Refinement Anomaly Detection (AFRAD), which integrates a multilayer perceptron-based stage-adaptive decoder that adaptively decodes multiscale feature maps to model broader contextual relationships. Furthermore, to minimize information loss, we introduce a convolution neural network-based focused local decoder to capture fine details at low-level dimensions and an MLP-based compensation decoder. The compensation decoder compensates for information missed by the stage-adaptive decoder and focused local decoder. This strategy improves the ability of the model to handle diverse aspects of the data, enabling robust anomaly detection. In addition, we experimentally demonstrate that the fusion of final representations enables the generation of high-quality

reconstructed feature maps. Our AFRAD achieves superior performance compared with conventional reconstruction-based methodologies on various public datasets.

Keywords: Anomaly detection, anomaly localization, multiscale feature refinement

1 Introduction

Anomaly detection is a critical task across various fields, with the aim of identifying patterns that deviate significantly from normal samples. Previous studies [1], [2], [3] mainly adopt a one-for-one approach, training a model on the normal data of a specific class for industrial. During this process, diverse techniques such as feature matching-based, score-based, and reconstruction-based methods, are employed to achieve notable improvements in detection accuracy and efficiency. However, diverse anomalies in industrial data, such as machine defects and electrical faults, limit the effectiveness of the one-for-one approach. In addition, obtaining anomalous images is often difficult, which limits the applicability of traditional methods. To address these limitations, recent studies have focused on multiclass unsupervised anomaly detection (MUAD). MUAD follows a one-for-all approach, allowing a single model to detect and localize anomalies across multiple classes without the need to distinguish between the normal data distributions of individual classes. Research on reconstruction-based approaches for MUAD has advanced based on the successful performances of convolutional neural networks-based (CNNs) and Transformer-based methods. These methods operate under the assumption that both normal and anomalous data can be reconstructed well, with anomaly detection and localization achieved by computing the difference between the input and reconstructed output. Normal data generally have low reconstruction errors; however, anomalous data have higher errors and can be analyzed to identify anomalies. Various approaches [4], [5], [6], [7], [8] including CNNs, Transformers, and diffusion models, have been explored to learn normal data distributions. However, each method presents significant challenges. CNNs struggle with long-range dependencies, making it difficult to capture global information owing to their limited ability to learn the relationships between elements that are far apart in a sequence. Beyond general industrial inspection, such anomaly detection and localization techniques are also highly relevant to information display applications, where fine-grained surface defects can directly affect visual quality, reliability, and production yield. In display manufacturing, defects such as scratches, contamination, and microstructural irregularities must be accurately identified for automated optical inspection and quality assurance. Recent work [9] has further shown that unsupervised anomaly segmentation is an important and practical direction for surface defect inspection in display panels.

To address these challenges, we propose Adaptive Feature Refinement Anomaly Detection (AFRAD), which is an innovative reconstruction-based MUAD framework that is designed to handle the complexity of anomaly detection. Although our experiments are conducted on widely used public industrial benchmarks for fair and

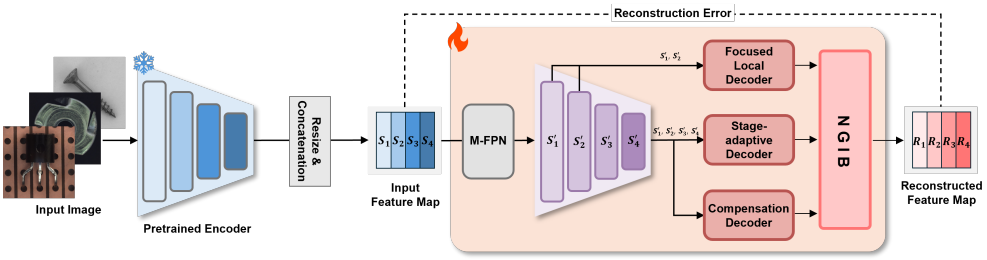


Fig. 1 Framework of AFRAD consisting of the M-FPN, focused local decoder, stage-adaptive decoder, compensation decoder, and NGIB. The input image is process by a pretrained CNN to extract multiscale features, which are then refined by the M-FPN. The refined multiscale features are decoded through three parallel branches: the focused local decoder, stage-adaptive decoder, and compensation decoder. These decoders capture different aspects of normality, and their outputs are integrated by the NGIB to produce the final reconstructed feature map. This reconstruction is used to compute anomaly scores for detection and localization.

reproducible evaluation, the proposed framework is also well aligned with defect localization scenarios arising in display inspection, where accurate multiscale reconstruction and localization are essential. The key components of AFRAD include a multiscale feature pyramid network (M-FPN), feature refinement module, and normality guided integration block (NGIB). The feature refinement module enhances multiscale feature representations, which are then processed by the focused local, stage-adaptive, and compensation decoders to effectively capture normality. The NGIB merges this optimized information into a detailed reconstructed token map. This comprehensive architecture effectively captures and integrates diverse information, ensuring robust performance in anomaly detection and localization. Our contributions are summarized as follows:

1. We designed the novel stage-adaptive decoder to enable the model to understand and process the intrinsic characteristics of complex information effectively.
2. We propose the M-FPN to improve multiscale feature interaction. It enables more effective integration of features at different scales, which is essential for handling defects of various sizes and types in a single image.
3. Our NGIB is built to integrate the rich information extracted from the previous scales, resulting in a highly detailed and context-aware reconstructed token map. This process allows the model to minimize information loss and improve its understanding of the object.
4. Our AFRAD achieves state-of-the-art (SoTA) performance while maintaining high computational efficiency. This demonstrates the ability of the framework to capture and synthesize important features for accurate anomaly detection.

2 Related Works

2.1 Unsupervised Anomaly Detection

Feature embedding-based methods have been studied in various forms [4], [8], [10], [11], [12], [13] by extracting features from normal images using pretrained models. RD4AD [8] uses WideResNet50 as a teacher model to extract features and employes a pair of teacher-student networks for feature reconstruction. Synthesis-based methods involve synthesizing abnormal images from normal images; various studies [6],[14],[15] have addressed these approaches. SimpleNet [14] added noise to the normal features in the feature space instead of generating random noise directly in the image. DeSTSeg [15] combines a pretrained teacher network, a segmentation network, and denoising student encoder-decoder into one integrated framework. Reconstruction-based methods use self-training encoders and decoders to reconstruct images for anomaly detection. Autoencoder (AE) models [7], [16], [17], [18], [19] are the most widely used reconstruction networks for anomaly detection. Specifically, DRAEM [7], which is a representative example of reconstruction-based techniques, enhances the generalization ability of the reconstruction network by introducing external datasets to synthesize abnormal images and reconstruct them as normal images. Transformer-based models, [20], [21], [22] have a high capacity to represent global information, suggesting that they can surpass AEs and become a new foundational reconstruction network for anomaly detection.

2.2 Multiclass Unsupervised Anomaly Detection

UniAD [22] proposed a unified reconstruction-based framework combining a pretrained encoder with a Transformer decoder for multiclass anomaly detection. OmniAL [23] was a unified CNN framework that trained the model using panel-guided synthetic anomaly data. DiAD [24] includes a diffusion-based framework for constructing pixel space AEs, latent space semantic-guided networks, and feature space pretrained feature extractors to maintain image categories and pixel-wise structural integrity across multiple classes.

3 Proposed Method

The pipeline of the proposed AFRAD is illustrated in Fig. 1. The input image is fed into a CNN-based pretrained model to extract and upsample multiscale feature maps, which are then concatenated and passed through the M-FPN. Finally, the reconstructed feature map is compared with the original feature map to generate an anomaly score map for detecting and localizing anomalies.

3.1 Feature Refinement Module

Conventional decoding approaches focus on refining local semantic details or modeling global contextual relationships. However, these approaches may leave critical information underexplored or overlooked. We propose a feature refinement module to address these challenges. As shown in Fig. 2, the feature refinement module refines

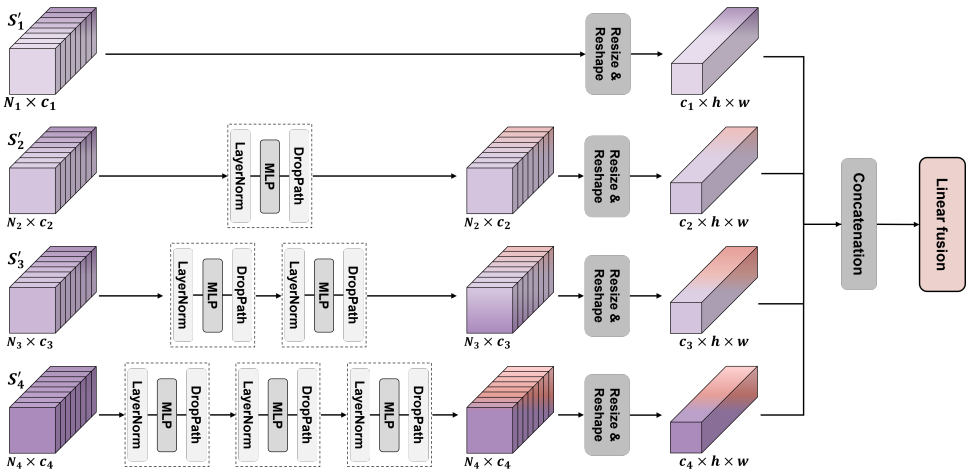


Fig. 2 Structure of the stage-adaptive decoder for rich feature representation. Stage-adaptive decoder adaptively enhances multiscale feature representations by controlling how many times the feature refinement module is applied at each stage. Refined feature representation from each stage are then resized and reshaped to a common resolution, and finally aggregated into a unified representation.

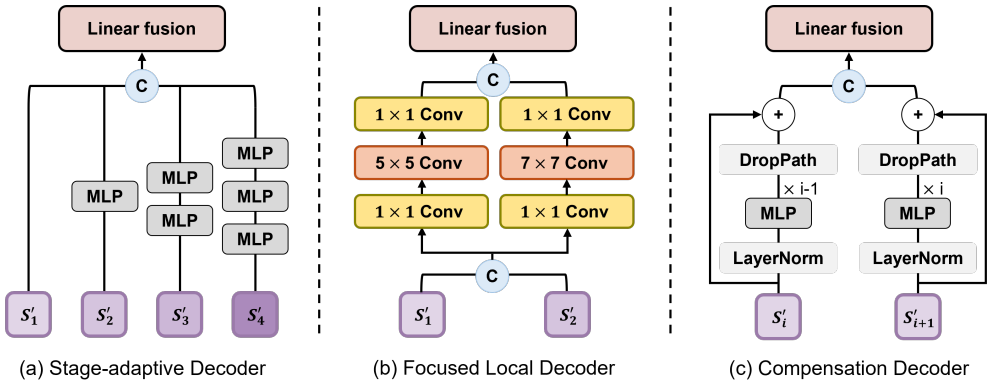


Fig. 3 (a) Structure of the stage-adaptive decoder for handling diverse information using MLP-based methods. (b) Structure of the focused local decoder for capturing local information using CNN-based methods. (c) Structure of the compensation decoder for addressing overlooked information from the stage-adaptive and focused local decoders by leveraging the feature refinement module to revisit and optimize multiscale feature representations.

and enhances multiscale feature representations before passing them to the decoder. The feature refinement module is implemented as a multilayer perceptron (MLP)-based method to enhance multiscale feature representations. The MLP is suitable for refining diverse feature representations without being tied to a specific structure. Furthermore, stacking multiple MLP layers transforms their inherently linear nature into a nonlinear expressive capacity through hierarchical processing. This significantly

enhances the ability of the model to represent complex patterns, making the feature refinement module a critical component for effectively addressing missed information and integrating multiscale features.

3.2 M-FPN

Industrial datasets often contain significantly fewer abnormal than normal data because abnormal samples are difficult to obtain. To address these challenges, we propose the M-FPN, which captures extensive information using patch embeddings of different kernel sizes on the features extracted from a pretrained backbone network. This process simulates the effect of processing multiple perspectives within a single image, thereby enhancing the ability of the model to capture diverse feature representations. We apply the feature refinement module to the patch-embedded features, reconstructing and refining the unique representations of each stage. This allows for effective processing of local semantics and global information before passing them to the decoder. We denote the channel, height, and width of the feature map for each stage. Mathematically, the M-FPN can be expressed as:

$$H_i' = \mathrm{PE}_i(S_i) \tag{1}$$

$$S_i' = \Phi_i(H_i') + H_i' \tag{2}$$

where S_i represents the input feature of the i-th layer, $S_i^{'}$ represents the final output feature of the i-th layer of M-FPN. PE represents patch embedding, and Φ represents the feature refinement module.

3.3 Stage-adaptive Decode

As shown in Fig. 3 (a), stage-adaptive decoder processes multiscale information by stacking MLP layers, enabling rich feature representation. This approach not only enhances the ability of the model to capture both fine-grained and broader contextual information but also enables it to handle complex tasks with a more comprehensive understanding of the input data. In particular, the high- dimensional pyramid levels represent global information such as the structural details of objects, whereas the lower-dimensional pyramid levels capture local information, including object details and semantic features. Rather than performing the same operation at each stage, we apply tailored operations that effectively process the unique information represented at each stage. In addition, to optimize the computational cost, we adjust the frequency of processing between the low and high-dimensional stages. Finally, to integrate each multiscale representation effectively, we pass them through a linear layer for fusion. From a computational perspective, focusing on local information in high dimensions and global information in low dimensions can avoid unnecessary details, allowing the fusion process to enhance the spatial relationships and model capacity to detect subtle anomalies accurately. Our stage-adaptive decoder can be expressed as follows:

$$\hat{S}_i = \mathrm{MLP}(S_i'), \quad \forall i \tag{3}$$

$$U_i = \mathrm{Upsample}(\hat{S}_i) \tag{4}$$

$$S_c = \text{Concat}(U_i), \quad \forall i \tag{5}$$

$$SAD_{\text{out}} = \text{Linear}(S_c) \tag{6}$$

The features $\hat{S_i} \in \mathbb{R}^{d_i \times h_i \times w_i}$ of each M-FPN stage are upsampled using bilinear interpolation, where $i \in \{1, 2, 3, 4\}$ denotes the index of the M-FPN stage. These upsampled features U_i are concatenated and passed to a linear layer for fusion. SAD_{out} represents the stage-adaptive decoder results.

3.4 Focused Local Decoder

As shown in Fig. 3 (b), the focused local decoder utilizes a convolution-based structure that focuses on capturing local information using only the first and second stages. Using all stages negatively affects both the performance and computational cost, whereas experiments have shown that focusing only on the first two stages is more effective for capturing low-level features. In our approach, stage 2 is upsampled to match the height and width of stage 1 using bilinear interpolation. The upsampled stage 2 is then concatenated with stage 1 and fed into the focused local decoder. The concatenated features are passed through a 1×1 convolution to capture the relationships between the feature maps containing different types of information. To capture local semantics while maintaining computational efficiency, we employ depth-wise convolutions with two different kernel sizes, following the approach used in SegNext [25]. This allows the model to focus on extracting detailed spatial information from each channel independently while reducing the computational burden compared with standard convolution. In this study, we select kernel sizes of 5 and 7 to capture different levels of detail across the feature maps. These local semantics are then concatenated and passed through a linear layer for fusion, thereby ensuring that the fine-grained details and various spatial features are integrated effectively integrated. The focused local decoder can be expressed as

$$\hat{S} = \text{Concat}(S'_1,\ \text{Upsample}(S'_2)) \tag{7}$$

$$\hat{S}_p = \text{Conv}_p\Big(\text{DW}_p\Big(\text{Conv}_p(\hat{S})\Big)\Big), \quad \forall p \tag{8}$$

$$S_c = \text{Concat}(\hat{S}_p), \quad \forall p \tag{9}$$

$$\text{FLD}_{\text{out}} = \text{Linear}(S_c) \tag{10}$$

where S'_1 and S'_2 represents the input stages 1 and 2, respectively. Conv represents the 1×1 convolution block and DW_p represents the depth-wise convolution block with kernel sizes p set to 5 and 7. FLD_{out} represents the results of the focused local decoder.

3.5 Compensation Decoder

As shown in Fig. 3 (c), the compensation decoder complements information missed by the stage-adaptive and focused local decoders, particularly intermediate-level and multiscale features that are often underrepresented. Specifically, it selectively processes two consecutive stages of multiscale features, allowing for a more targeted refinement of information that may have been overlooked. The compensation decoder addresses

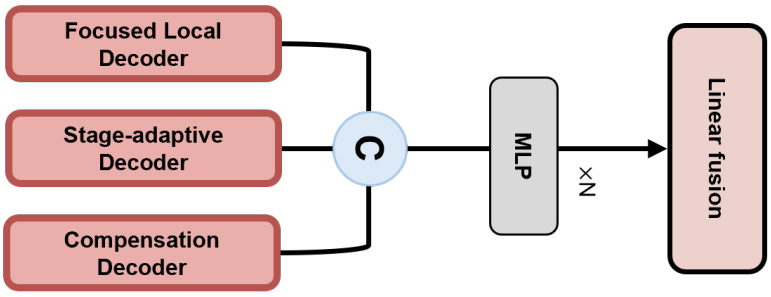


Fig. 4 Structure of the NGIB that integrates high-quality feature maps reconstructed by the stage-adaptive, focused local, and compensation decoders.

these limitations using the feature refinement module, which revisits and optimizes the selected multiscale feature representations before passing through the decoder, ensuring richer and more refined outputs. As shown in Table 6, the experimental results demonstrate that the optimization settings of the compensation decoder and its integration with multiscale feature maps significantly enhance the anomaly detection and localization performance by revisiting and refining multiscale feature representations, which compensate for the missing details and contextual gaps left by the focused local and stage-adaptive decoders.

3.6 NGIB

We introduce the NGIB to fuse the diverse information captured by the stage-adaptive, focused local, and compensation decoders. As shown in Fig. 4, NGIB is designed to integrate the high-quality reconstructed feature maps by leveraging an MLP, similar to the stage-adaptive decoder, to prioritize the most semantically important features within each normality aspect. By utilizing linear layers to integrate this information efficiently, the NGIB not only combines the features but also preserves the semantic consistency of the data. This process allows the model to develop a refined and context-aware representation of the input data, effectively balancing fine-grained local details with the broader global context. This comprehensive approach significantly enhances the accuracy of anomaly detection and localization by enabling the model to identify subtle anomalies that may be overlooked by less integrated methods. The NGIB functions as an advanced fusion mechanism, surpassing simple feature aggregation by facilitating the semantically guided integration of information. The fusion can be expressed as follows:

$$F_c = \text{Concat}(SAD_{\text{out}},\ CD_{\text{out}},\ FLD_{\text{out}}) \tag{11}$$

$$\text{FB}_{\text{out}} = \text{MLP}(F_c) \tag{12}$$

$$F_{\text{rec}} = \text{Linear}(\text{FB}_{\text{out}}) \tag{13}$$

Table 1 Comparison with state-of-the-art methods on MVTec-AD dataset for multi-class anomaly detection with I-AUROC/P-AUROC/P-AUPRO metrics. All methods are evaluated under the unified case. Bold and underlining indicate best results and second-best results, respectively.

Method		Non-Reconstruction Method			Reconstruction-based Method			
Category		RD4AD	SimpleNet	DeSTSeg	DRAEM	UniAD	DiAD	Ours
Objects	Bottle	99.6/97.8/**94.0**	**100.**/97.2/89.0	98.7/93.3/67.5	97.5/87.6/80.7	99.7/98.1/93.1	99.7/**98.4**/86.6	**100.**/97.9/93.5
	Cable	84.1/85.1/75.1	97.5/96.7/85.4	89.5/89.3/49.4	57.8/71.3/40.1	95.2/**97.3**/86.1	94.8/96.8/80.5	**98.8**/96.5/**89.8**
	Capsule	94.1/**98.8**/**94.8**	90.7/98.5/84.5	82.8/95.8/62.1	65.3/50.5/27.3	86.9/98.5/92.1	89.0/97.1/87.2	**95.2**/**98.8**/93.8
	Hazelnut	60.8/97.9/92.7	99.9/**98.4**/87.4	98.8/98.2/84.5	93.7/96.9/78.7	99.8/98.1/**94.1**	99.5/98.3/91.5	**100.**/98.1/94.0
	Metal Nut	**100.**/94.8/**91.9**	96.9/**98.0**/85.2	92.9/84.2/53.0	72.8/62.2/66.4	99.2/94.8/81.8	99.1/97.3/90.6	99.2/96.2/90.1
	Pill	97.5/**97.5**/**95.8**	88.2/96.5/81.9	77.1/96.2/27.9	82.2/94.4/53.9	93.7/95.0/95.3	95.7/95.7/89.0	97.7/**97.0**/95.1
	Screw	97.7/**99.4**/**96.8**	76.7/96.5/84.0	69.9/93.8/47.3	92.0/95.5/55.2	87.5/98.3/95.2	**99.7**/97.9/95.0	97.4/**99.4**/96.4
	Toothbrush	97.2/**99.0**/92.0	89.7/98.4/87.4	71.7/96.2/30.9	90.6/97.7/68.9	94.2/98.4/87.9	**99.7**/**99.0**/**95.0**	90.8/98.3/88.9
	Transistor	94.2/85.9/74.7	99.2/95.8/83.2	78.2/73.6/43.9	74.8/64.5/39.0	**99.8**/**97.9**/**93.5**	**99.8**/95.1/90.0	98.8/94.3/89.0
	Zipper	**99.5**/**98.5**/94.1	99.0/97.9/90.7	88.4/97.3/66.9	98.8/98.3/91.9	95.8/96.8/92.6	95.1/96.2/91.6	98.3/98.2/**94.3**
Textures	Carpet	98.5/**99.0**/**95.1**	95.7/97.4/90.6	95.9/93.6/89.3	98.0/98.6/93.1	99.8/98.5/94.4	99.4/98.6/90.6	**99.9**/98.4/94.4
	Grid	98.0/96.5/**97.0**	97.6/96.8/88.6	97.9/97.0/86.8	**99.3**/**98.7**/92.1	98.2/96.5/92.9	98.5/96.6/94.0	99.1/97.4/92.0
	Leather	**100.**/99.3/**97.4**	**100.**/98.7/92.7	99.2/**99.5**/91.1	98.7/97.3/88.5	**100.**/98.8/96.8	99.9/98.8/91.3	**100.**/98.2/95.9
	Tile	98.3/95.3/85.8	99.3/95.7/90.6	97.0/93.0/87.1	99.8/**98.0**/**97.0**	99.3/91.8/78.4	96.8/92.4/90.7	**100.**/91.7/78.1
	Wood	99.2/95.3/90.0	98.4/91.4/76.3	**99.9**/95.9/83.4	99.8/**96.0**/94.2	98.6/93.2/86.7	99.7/93.3/**97.5**	98.9/92.4/84.8
Mean		94.6/96.1/91.1	95.3/**96.9**/86.5	89.2/93.1/64.8	88.1/87.2/71.1	96.5/96.8/90.7	97.2/96.8/90.7	**98.3**/**96.9**/**91.3**

where CD_{out} denotes the compensation decoder results. Finally, the fused feature map F_{rec} is used to score the anomaly.

3.7 Anomaly Detection and Localization

During the training phase, we train the model using the MSE loss between F_{rec} and F_{ori}, where F_{rec} is the feature map fused by the fusion block and F_{ori} is the concatenated multiscale feature map extracted from the pretrained model. During the inference phase, the reconstructed feature map obtained through the aforementioned processes is compared with the feature map extracted using the pretrained model. The difference between these feature maps is calculated using the L2 norm to generate the anomaly score map M, as follows: M as follows :

$$\mathrm{M} = \sqrt{\sum (F_{\text{rec}} - F_{\text{ori}})^2}. \tag{14}$$

For detection, the anomaly score map is evaluated to determine the presence of anomalies using average pooling and max operations. For localization, the anomaly score map, in which each pixel is assigned an anomaly score, is computed as the L2 norm of the reconstruction differences. The resulting values are then upsampled using bilinear interpolation to produce localization results at the original image resolution.

4 Experiments

4.1 Datasets and Metrics

The MVTec-AD [26] dataset is a comprehensive industrial anomaly detection dataset with 15 classes, that simulates real-world production scenarios. It includes 5,354 high-resolution images across 5 textures and 10 objects. The training set contains 3,629 anomaly-free images, whereas the test set contains 1,725 images of both normal and abnormal samples, including pixel-level annotations for anomaly localization. The VisA [27] dataset consists of 10,821 high-resolution images, of which 9,621 are normal and 1,200 are anomalous images, containing 78 types of anomalies. The dataset

Table 2 Comparison with state-of-the-art methods on VisA dataset for multi-class anomaly detection with I-AUROC/P-AUROC/P-AUPRO metrics. All methods are evaluated under the unified case. Bold and underlining indicate best results and second-best results, respectively.

Method Category		DRAEM	SimpleNet	DeSTSeg	UniAD	DiAD	Ours
Complex structure	pcb1	83.9/94.0/52.9	91.6/99.2/83.6	87.6/95.8/83.2	92.8/93.3/64.1	88.1/98.7/80.2	**95.4**/**99.4**/**91.1**
	pcb2	81.7/94.1/66.2	**92.4**/96.6/**85.7**	86.5/97.3/79.9	87.8/93.9/66.9	91.4/95.2/67.0	92.2/**98.1**/84.9
	pcb3	87.7/94.1/43.0	89.1/97.2/**85.1**	**93.7**/97.7/62.4	78.6/97.3/70.6	86.2/96.7/68.9	88.0/**98.7**/84.7
	pcb4	87.1/72.3/75.7	97.0/93.9/61.1	97.8/95.8/76.9	98.8/94.9/72.3	**99.6**/97.0/**85.0**	99.4/**97.7**/84.0
Multiple instances	macaroni1	68.6/89.8/67.0	85.9/98.9/92.0	76.6/99.1/62.4	79.9/97.4/84.0	85.7/94.1/68.5	**90.1**/**99.4**/**96.5**
	macaroni2	60.3/83.2/65.3	68.3/93.2/77.8	68.9/**98.5**/70.0	71.6/95.2/76.6	62.5/93.6/73.1	**82.1**/97.7/**88.6**
	capsules	**89.6**/96.6/62.9	74.1/97.1/73.7	87.1/96.9/76.7	55.6/88.7/43.7	58.2/97.3/**77.9**	67.6/**98.5**/75.0
	candle	70.2/82.6/65.6	84.1/97.6/87.6	94.9/98.7/69.0	94.1/98.5/91.6	92.8/97.3/89.4	**97.0**/**99.3**/**94.9**
Single instance	cashew	67.3/68.5/38.5	88.0/**98.9**/84.1	92.0/87.9/66.3	**92.8**/98.6/87.9	91.5/90.9/61.8	91.4/98.3/**89.0**
	chewinggum	90.0/92.7/41.0	96.4/97.9/78.3	95.8/98.8/68.3	96.3/98.8/81.3	99.1/94.7/59.5	**99.2**/**99.4**/**87.1**
	fryum	86.2/83.2/69.5	88.4/93.0/**85.1**	**92.1**/88.1/47.7	83.0/95.9/76.2	89.8/**97.6**/81.3	87.7/97.1/82.1
	pipe fryum	87.1/72.3/61.9	90.8/98.5/83.0	94.1/98.9/45.9	94.7/98.9/91.5	96.2/**99.4**/89.9	**97.0**/98.9**94.3**
Mean		80.5/87.0/59.1	87.2/96.8/81.4	88.9/96.1/67.4	85.5/95.9/75.6	86.8/96.0/75.2	**90.6**/**98.5**/**87.7**

is divided into 12 subsets, each corresponding to a distinct object, which are categorized into complex structure, multiple instances, and single instance. The CIFAR-10 [28] dataset is an image classification dataset with 10 classes. We used the area under the receiver operator curve (AUROC) as the evaluation metric for anomaly detection. The AUROC indicates how well a model distinguishes between normal and anomalous data at both the image and pixel levels. We denote the image and pixel-level AUROC as I-AUROC and P-AUROC, respectively. Additionally, we report pixel-level anomaly localization by the pixel-level area under the per-region overlap (P-AUPRO).

4.2 Implementation Details

In the experiments, all images in the MVTec-AD, VisA and CIFAR-10 dataset were resized to 256 $\times$ 256 pixels. We adopted the pretrained EfficientNet-B4 as the feature extractor. The AdamW optimizer was employed with a learning rate of 0.0001. The model was trained for 1000 epochs on the MVTec-AD and VisA datasets and 200 epochs on the CIFAR- 10 dataset, utilizing an NVIDIA RTX 4090 GPU. The loss function used during training was the sum of the MSE calculated across various scales.

4.3 Comparisons with State-of-the-art methods

We conducted experiments to compare our method with SoTA methods on the MVTec-AD, VisA, and CIFAR-10 datasets.

4.3.1 MVTec-AD dataset

We categorized the methods into reconstruction-based and non-reconstruction methods for comparison. For non-reconstruction methods, we selected embedding-based methods such as Padim [29], MKD [30], DeSTSeg [15], and RD4AD [8], as well as the synthesis-based method SimpleNet [14]. For reconstruction-based methods, we selected UniAD [22] and DRAEM [7]. In addition, we included DiAD [24], which is a diffusion-based reconstruction method, for comparison.

As shown in Table 1, our AFRAD outperformed several SoTA approaches on the MVTec-AD dataset for MUAD, as evidenced by its superior performance in both the I-AUROC and P-AUROC metrics. Overall, the mean I- AUROC/P-AUROC scores

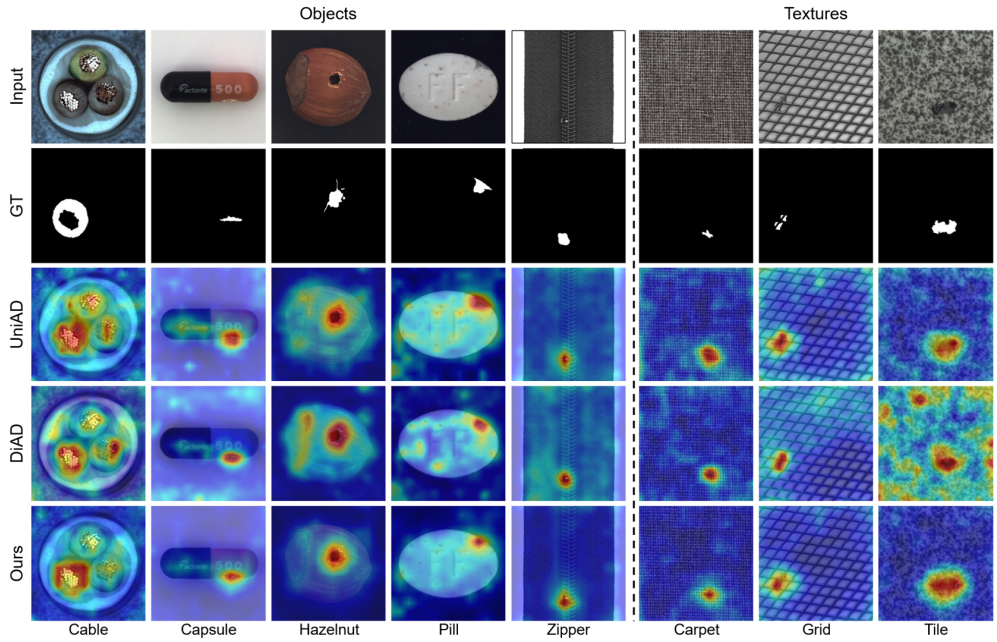


Fig. 5 Qualitative results for anomaly localization on MVTec-AD dataset. On the left side of the dotted line are object categories and on the right side are texture categories. From top to bottom: sample input images; ground truths; and the predicted anomaly heatmaps of UniAD, DiAD and our predicted anomaly heatmaps.

Table 3 Anomaly detection results with AUROC metric on CIFAR-10. In this context, "01234" represents the samples from classes 0, 1, 2, 3, and 4 that are considered as the normal ones. All methods are evaluated under the unified case. Bold and underlining indicate best results and second-best results, respectively

Method Normal Indices	US	FCDD	FCDD+OE	PANDA	MKD	UniAD	Ours
01234	51.3	55.0	71.8	66.6	64.2	<u>84.4</u>	**84.8**
56789	51.3	50.3	73.7	73.2	69.3	<u>80.9</u>	**81.2**
02468	63.9	59.2	85.3	77.1	76.4	<u>93.0</u>	**93.4**
13579	56.8	58.5	85.0	72.9	78.7	**90.6**	<u>90.2</u>
Mean	55.9	55.8	78.9	72.4	72.1	<u>87.2</u>	**87.4**

for our method were 98.4/96.9, which were the highest among all compared methods, indicating the robustness and effectiveness of AFRAD in detecting and localizing anomalies across the different categories. As illustrated in Fig. 5, our model achieved significant improvements in anomaly localization compared with UniAD and DiAD. The heatmaps generated from the three models clearly show that our model targeted the anomaly regions more accurately. Specifically, our model produced higher intensity around the anomalous regions, while maintaining lower levels in the background and

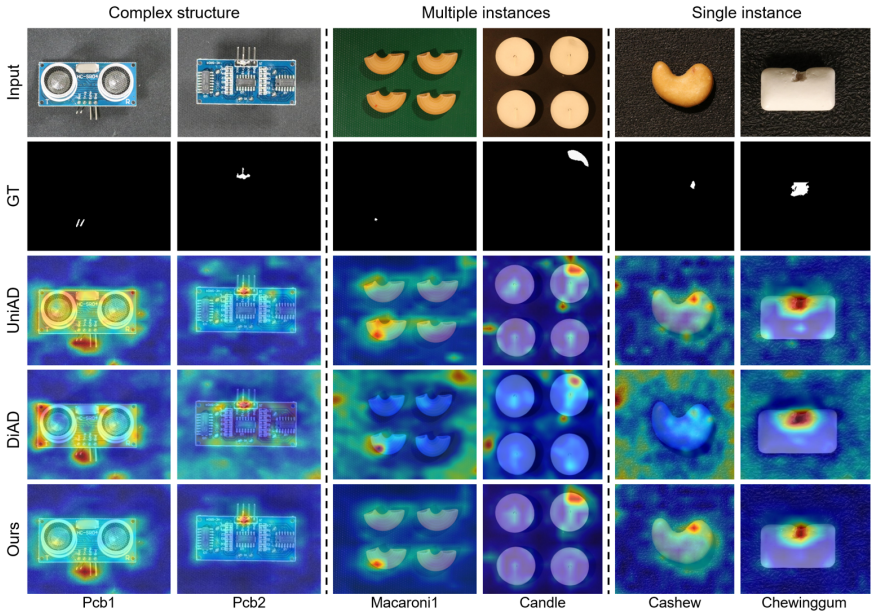


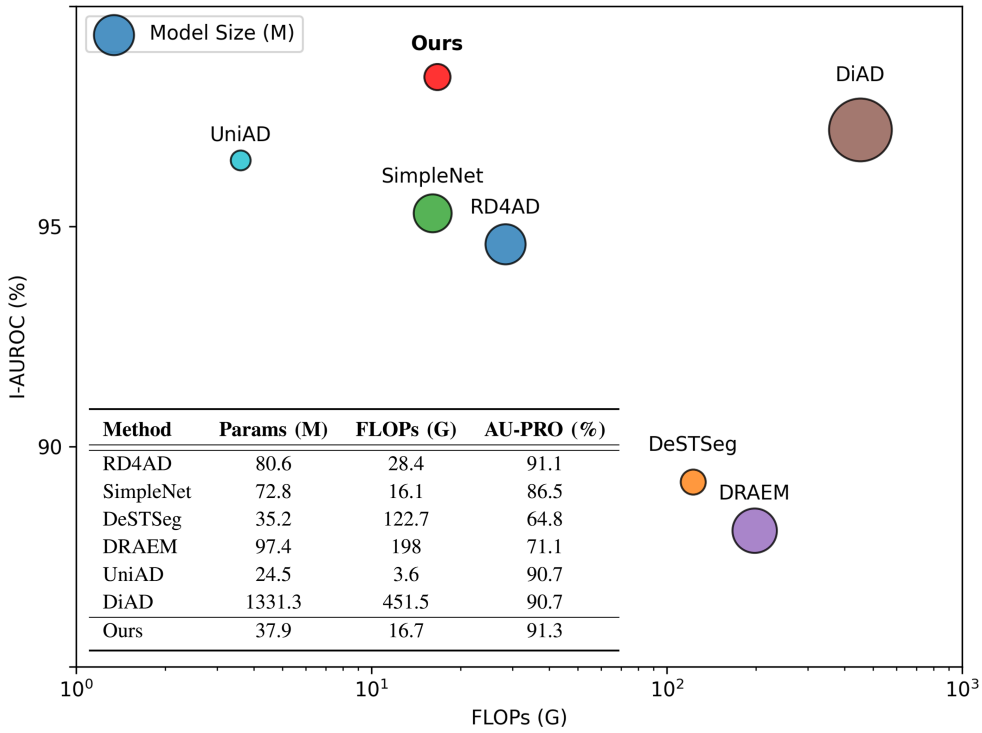
Fig. 6 Qualitative results for anomaly localization on VisA dataset. On the left side of the dotted line are complex structure categories and on the middle side are multiple instances categories and on the right side are single instance categories. From top to bottom: sample input images; ground truths; and the predicted anomaly heatmaps of UniAD, DiAD and our predicted anomaly heatmaps.

other unrelated regions. This difference highlights the enhanced capability of our model to localize anomalies accurately without being influenced by the surrounding noise, resulting in clearer and more reliable detection outcomes.

4.3.2 VisA dataset

We compared our method with RD4AD, SimpleNet, DeSTSeg, UniAD, and DiAD. Table 2 presents the experimental results for the VisA dataset, which is more complex and challenging than the simpler MVTec-AD dataset. AFRAD achieved the highest I-AUROC/P-AUROC scores, not only in complex structures but also in both the multiple and single-instance categories. These results demonstrate the robustness of the proposed method in addressing a wide range of defects.

As illustrated in Fig. 6, our model exhibited significant improvements in anomaly localization compared with existing models, not only in simple structures but also in complex and challenging structures. Our model accurately targeted anomalous regions in single as well as multiple instances. Specifically, our model remained robust to the background and other unrelated regions. This difference highlights the enhanced capability of our model to localize anomalies accurately without being influenced by the surrounding noise, resulting in clearer and more reliable detection outcomes.



Method	Params (M)	FLOPs (G)	AU-PRO (%)
RD4AD	80.6	28.4	91.1
SimpleNet	72.8	16.1	86.5
DeSTSeg	35.2	122.7	64.8
DRAEM	97.4	198	71.1
UniAD	24.5	3.6	90.7
DiAD	1331.3	451.5	90.7
Ours	37.9	16.7	91.3

Fig. 7 Comparison of computational efficiency between AFRAD and SoTA methods. AFRAD achieves high I-AUROC while maintaining low computational cost in terms of GFLOPs and parameter size, demonstrating a favorable trade-off between performance and efficiency.

4.3.3 CIFAR-10 dataset

To evaluate the effectiveness of our AFRAD further, we adapted CIFAR-10 to a unified scenario involving four combinations. For each combination, five categories were designated as normal samples, and the remaining categories were considered anomalies. The class the indices for the four combinations were 01234, 56789, 02468, and 13579, respectively. As shown in Table 3, our AFRAD achieved the highest performance, with a score of 87.4, surpassing US [4], FCDD [31], FCDD-OE [31] (which uses CIFAR-100 as outlier exposure), PANDA [32], MKD, and UniAD.

4.4 Computational efficiency

Considering computational efficiency is important for real-world applications. Therefore, we compared our method with SoTA models, as shown in Fig. 7. AFRAD achieved high I-AUROC with low GFLOPs and parameter size. Specifically, AFRAD outperforms several competing methods in terms of efficiency by achieving a favorable balance between computational cost and detection accuracy. The comparison was conducted under identical settings, with the batch size set to 1 and the input image resolution fixed at 256 size. Table 4 reports the wall-clock inference time of different anomaly

Table 4 Wall-clock inference time comparison of different anomaly detection methods. All measurements are reported in milliseconds per image (ms/image) under the same experimental setting. Lower is better.

Method	**Inference Time (ms/image)**
DeSTSeg	9.4
SimpleNet	13.0
DRAEM	22.2
Ours	24.5
RD4AD	41.0
HVQ-Trans	50.0
UniAD	85.0
EfficientAD	125.0
PatchCore	148.0
ReconPatch	150.0
Uniformaly	200.0
DiAD	675.0
DDAD	1192.0
AnoDDPM	11950.0

detection methods in milliseconds per image (ms/image) under the same experimental setting. Our method requires 24.5 ms/image, which is substantially faster than several recent reconstruction-based and diffusion-based approaches, including UniAD (85 ms), DiAD (675 ms), DDAD (1192 ms), and AnoDDPM (11950 ms). These results indicate that the proposed AFRAD achieves favorable practical efficiency while maintaining strong anomaly detection performance. Although a few lightweight methods such as DeSTSeg and SimpleNet show lower latency, our method still provides a competitive inference speed and offers a balanced trade-off between detection performance and computational cost.

Table 5 Results of ablation experiments on different decoders and NGIB. Results are reported using I-AUROC, P-AUROC and P-AUPRO.

Focused Local Decoder	Compensation Decoder	Stage-adaptive Decoder	NGIB	Mean
-	✓	✓	✓	91.7/95.1/87.6
✓	-	✓	✓	91.6/94.6/87.4
✓	✓	-	✓	91.2/94.7/87.6
✓	✓	✓	-	90.9/94.8/87.8
✓	✓	✓	✓	**98.3/96.9/91.3**

Table 6 Ablation experiments on stage configurations for each decoder. Results are reported using I-AUROC and P-AUROC. Here, S refers to the multi-scale stages.

Focused Local Decoder		Compensation Decoder		Stage-adaptive Decoder		Mean
$S_{1,2}$	$S_{1,2,3,4}$	$S_{1,2}$	$S_{2,3}$	$S_{3,4}$	$S_{1,2,3,4}$	
✓	-	✓	-	✓	-	96.4/96.5
✓	-	-	✓	✓	-	96.4/96.5
✓	-	-	✓	-	✓	95.9/96.6
-	✓	✓	-	-	✓	95.6/96.5
✓	-	✓	-	-	✓	**98.3/96.9**

4.5 Ablation study

4.5.1 Effectiveness of Each Component

As shown in Table 5, we conducted experiments by selectively enabling or disabling different decoders and NGIB. When excluding NGIB, the performance dropped significantly even when multiple aspects of normality were processed through different decoders. without an effective fusion mechanism the benefit of decoding diverse normality features is limited, leading to the lowest performance among the configurations. The full configuration with all components enabled yielded the best performance, confirming their synergy in improving anomaly detection and localization.

4.5.2 Comparison of the effectiveness of decoder

As shown in Table 6, we conducted experiments by varying the stage inputs to each decoder. The results show that using all four multiscale stages $S_{1,2,3,4}$ in the stage-adaptive decoder consistently improved performance. For the focused local decoder and the compensation decoder, the best results were obtained by using $S_{1,2}$. Although these two decoders operate on similar low-level features, they employ different processing mechanisms, i.e., convolution-based local refinement and MLP-based compensation. The superior performance obtained when both branches are jointly used indicates that they capture complementary aspects of normality rather than redundant information. These findings suggest that high-level global structure is more effectively modeled by the stage-adaptive decoder, whereas low-level spatial details and fine appearance patterns are better preserved by the focused local and compensation decoders. This observation supports that the proposed multi-branch design is not a simple aggregation of modules, but a functionally differentiated architecture for learning normality across feature scales.

4.5.3 Effectiveness of FRM

As shown in Table 7, we analyzed the effect of the Feature Refinement Module (FRM) on model performance. Without FRM, the performance on MVTecAD and VisA was 89.0/93.9 and 79.2/95.8, respectively. When FRM was incorporated, the performance

Table 7 Ablation study on the effect of FRM. Results are reported using I-AUROC and P-AUROC.

FRM	MVTecAD	VisA
-	89.0 / 93.9	79.2 / 95.8
✓	**98.3 / 96.9**	**83.7 / 96.6**

substantially improved to 98.3/96.9 and 83.7/96.6. This demonstrates that FRM effectively integrates multi-scale information, thereby enhancing detection capability at both the image-level and pixel-level. In particular, the preservation of multi-scale features proves essential for detecting small defects and texture-based anomalies. These results experimentally validate that FRM is a crucial component of the proposed framework.

5 Conclusion

This study proposes a novel reconstruction-based framework, AFRAD, for multiclass unsupervised anomaly detection in industrial domains, including manufacturing, quality control, and predictive maintenance. The experimental results demonstrated that our AFRAD outperformed SoTA methods on datasets such as MVTec-AD, VisA and CIFAR-10, achieving high accuracy at both the image and pixel levels. This highlights its robust performance, even in cases of data imbalance and various anomaly patterns that commonly occur in industrial datasets. Although our AFRAD sets a new standard for addressing complex anomaly detection challenges, several limitations remain. AFRAD may require additional domain adaptation strategies when applied to datasets with significant domain shifts. Future work will explore domain adaptation to improve the flexibility and integrate few-shot or zero-shot learning to reduce the dependency on extensive training data while maintaining high performance.

Acknowledgements. This research was supported by LG Electronics Co., Ltd., the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT) (IITP-2026-RS-2023-00260091), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-16066849 and RS-2024-00414230), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25399644).

References

[1] Ma, X., Keung, J., He, P., Xiao, Y., Yu, X., Li, Y.: A semisupervised approach for industrial anomaly detection via self-adaptive clustering. IEEE Transactions on Industrial Informatics **20**(2), 1687–1697 (2023)

[2] Qiao, Y., Lü, J., Wang, T., Liu, K., Zhang, B., Snoussi, H.: A multihead attention self-supervised representation model for industrial sensors anomaly detection. IEEE Transactions on Industrial Informatics **20**(2), 2190–2199 (2023)

[3] Xu, Q., Xie, T., Jiang, C., Cheng, Q., Wang, X.: Adaptive working condition recognition with clustering-based contrastive learning for unsupervised anomaly detection. IEEE Transactions on Industrial Informatics (2024)

[4] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4183–4192 (2020)

[5] Yi, J., Yoon, S.: Patch svdd: Patch-level svdd for anomaly detection and segmentation. In: Proceedings of the Asian Conference on Computer Vision (2020)

[6] Li, C.-L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9664–9674 (2021)

[7] Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8330–8339 (2021)

[8] Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9737–9746 (2022)

[9] Song, J.-W., Kong, K., Park, Y.-I., Kim, S.-G., Kang, S.-J.: Self-supervised anomaly segmentation for surface defect inspection in display panels. Journal of the Society for Information Display **33**(11), 1059–1067 (2025)

[10] Shao, H., Peng, J., Shao, M., Liu, B.: Multi-scale prototype fusion network for industrial product surface anomaly detection and localization. IEEE Sensors Journal (2024)

[11] Sohn, K., Li, C.-L., Yoon, J., Jin, M., Pfister, T.: Learning and evaluating representations for deep one-class classification. arXiv preprint arXiv:2011.02578 (2020)

[12] Wang, G., Han, S., Ding, E., Huang, D.: Student-teacher feature pyramid matching for anomaly detection. arXiv preprint arXiv:2103.04257 (2021)

[13] Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328

(2022)

[14] Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20402–20411 (2023)

[15] Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: Destseg: Segmentation guided denoising student-teacher for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3914–3923 (2023)

[16] Gao, Y., Han, Z., Wang, J.: Effd: An unsupervised surface defect detection method based on estimation and fusion of normal sample feature distribution. IEEE Sensors Journal (2024)

[17] Yan, Y., Wang, D., Zhou, G., Chen, Q.: Unsupervised anomaly segmentation via multilevel image reconstruction and adaptive attention-level transition. IEEE Transactions on Instrumentation and Measurement **70**, 1–12 (2021)

[18] Huang, W., Liu, Z., Jin, X., Xu, J., Yao, X.: Improved autoencoder model with memory module for anomaly detection. IEEE Sensors Journal (2024)

[19] Zavrtanik, V., Kristan, M., Skočaj, D.: Dsr–a dual subspace re-projection network for surface anomaly detection. In: European Conference on Computer Vision, pp. 539–554 (2022). Springer

[20] You, Z., Yang, K., Luo, W., Cui, L., Zheng, Y., Le, X.: Adtr: Anomaly detection transformer with feature reconstruction. In: International Conference on Neural Information Processing, pp. 298–310 (2022). Springer

[21] Pirnay, J., Chai, K.: Inpainting transformer for anomaly detection. In: International Conference on Image Analysis and Processing, pp. 394–406 (2022). Springer

[22] You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., Le, X.: A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems **35**, 4571–4584 (2022)

[23] Zhao, Y.: Omnial: A unified cnn framework for unsupervised anomaly localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3924–3933 (2023)

[24] He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., Xie, L.: A diffusion-based framework for multi-class anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 8472–8480 (2024)

[25] Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., Hu, S.-M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in Neural Information Processing Systems **35**, 1140–1156 (2022)

[26] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9592–9600 (2019)

[27] Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European Conference on Computer Vision, pp. 392–408 (2022). Springer

[28] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

[29] Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: International Conference on Pattern Recognition, pp. 475–489 (2021). Springer

[30] Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M.H., Rabiee, H.R.: Multiresolution knowledge distillation for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14902–14912 (2021)

[31] Liznerski, P., Ruff, L., Vandermeulen, R.A., Franks, B.J., Kloft, M., Müller, K.-R.: Explainable deep one-class classification. arXiv preprint arXiv:2007.01760 (2020)

[32] Reiss, T., Cohen, N., Bergman, L., Hoshen, Y.: Panda: Adapting pretrained features for anomaly detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2806–2814 (2021)