International Journal on Semantic Web and Information Systems
Volume 20 • Issue 1

Semantic Coarse-to-Fine Granularity Learning for Two-Stage Few-Shot Anomaly Detection

Lei Zhang, Hubei University of Automotive Technology, China
Chengzhi Lyu, Hubei University of Automotive Technology, China*
Ziheng Chen, Hubei University of Automotive Technology, China
Shaokang Li, Hubei University of Automotive Technology, China
Bin Xia, Hubei University of Automotive Technology, China

ABSTRACT

Anomaly detection is critical in industrial inspection, where identifying defects significantly impacts product quality and safety. Existing models, primarily based on convolutional neural networks (CNNs), struggle with noise sensitivity and insufficient resolution for fine-grained feature discrimination. To address these issues, we propose a two-stage few-shot anomaly detection network that enhances semantic feature granularity and generalization. The network includes a coarse-grained anomaly detection module, a multi-scale channel attention module, and a fine-grained detection module. The coarse-grained module identifies abnormal regions, serving as the initial filter. The multi-scale channel attention module focuses on anomalous features, enhancing sensitivity to fine-grained characteristics. This step overcomes limitations in discerning subtle yet critical anomalies. The fine-grained detection module refines feature maps, enhancing generalization. Experimental results on the MVTec dataset show an image-level Area under the region of convergence (AUROC) of 92.3% and a pixel-level AUROC of 95.3%, a 1% to 2% improvement over leading FSAD methods.

KEYWORDS

Anomaly Detection, Few-Shot Learning, Neural Networks, Unsupervised Learning

INTRODUCTION

Anomaly detection plays a crucial role in industrial settings by identifying irregularities such as scratches and broken parts, thereby enhancing production efficiency. While obtaining normal samples is typically straightforward in these environments, acquiring diverse and challenging defective examples can be challenging. The complexity of the process often renders classical supervised training (Ding et al., 2022; Li et al., 2023; Liu et al., 2023; Bozorgtabar & Mahapatra, 2023) impractical, leading to the prevalence of unsupervised methods in industrial anomaly detection. These methods, which only use normal samples during training, identify anomalies by contrasting the tested data with

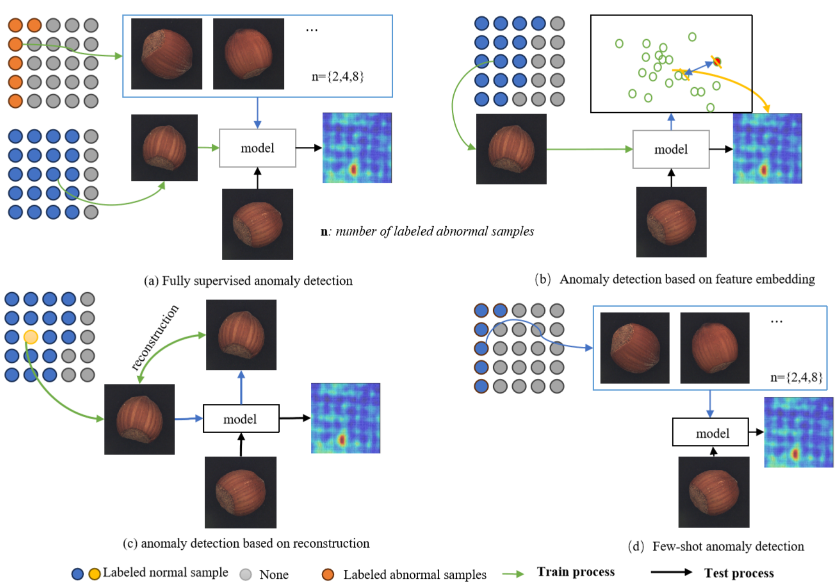
DOI: 10.4018/IJSWIS.344426 *Corresponding Author

learned normal features (Ilyas et al., 2022; Salehi et al., 2021; Xu et al., 2023). While unsupervised anomaly detection methods (Sun et al., 2023; Fang et al., 2023) primarily focus on feature learning to capture normal data's intrinsic characteristics, recent approaches allow for the labeling of a small number of anomalous samples (Atabay & Hassanpour, 2023), albeit at an increased cost.

To address the challenges of limited sample images and reduce labeling costs, few-shot anomaly detection has been proposed. Conventional supervised anomaly detection relies on a combination of limited anomaly data and a large number of normal samples to detect anomalies (Ding et al., 2022; Atabay & Hassanpour, 2023; Bozorgtabar & Mahapatra, 2023; Pang, Yan, et al.,2020), as shown in Figure 1(a), but it often exhibits inferior performance compared to unsupervised methods in anomaly identification and localization. In contrast, embedding-based, unsupervised anomaly detection methods leverage pre-trained models (Wang et al. 2023) eliminating the need for a large amount of training data, as shown in Figure 1(b). On the other hand, unsupervised anomaly detection methods based on image reconstruction require training the reconstruction model from scratch, necessitating a larger training set, as shown in Figure 1(c). However, both methods still require adjustments to fit unseen categories.

Recent studies focus on few-shot anomaly detection. The aim of generalized few-shot anomaly detection is to use a limited number of labeled anomalies as partial knowledge of anomalies within a specific domain of interest for training (Sheynin et al., 2021) requiring only a small number of samples for each category, as illustrated in Figure 1(d). In the early stages, the form of transfer learning was used to improve the learning effect in related fields, utilizing knowledge from the source domain to assist in semantic anomaly detection in the target domain. To address the problem caused by insufficient abnormal samples, a single model is used for detection across multiple categories, and fine-tuning is conducted based on a small number of high-quality samples. Adversarial models are used for sample generation, and multi-scale convolutional networks are combined to differentiate images, thereby greatly reducing the demand for training samples. However, the adjustability of the adversarial model may pose challenges, and its generalization may decrease with an increase in the number of training samples.

Figure 1. Four Different Common Anomaly Detection Methods



To enhance the model's generalizability, we propose a novel framework termed the two-stage, few-shot anomaly detection model. This model is a deep, one-class anomaly detection algorithm designed to explore fine-grained feature distributions for the few-shot anomaly detection problem (Ristea et al., 2022). Drawing inspiration from the human cognitive process of comparing differences between two objects, the algorithm brings anomaly detection closer to real-world performance. Training often stagnates when samples are aggregated for training on new products in production lines. For the problem of frequently needing to switch scenarios, few-shot anomaly detection is more suitable for this challenge.

The study proposes a two-stage, few-shot anomaly detection approach that aims to learn the robust feature distribution of images captured under conditions of clean and consistent lighting. The network architecture is easy to understand and incorporates the widely used residual neural network (ResNet). Despite its simplicity, it demonstrates outstanding performance and yields improved results, enhancing the model's capability to comprehend location information. In summary, the contributions of this paper can be summarized as follows:

- To make the features represent a more fine-grained distribution, we propose a few-shot anomaly detection framework from coarse to fine, dividing the coarse and fine anomalies into two stages;
- Multiple categories are aggregated to extract a common framework from normal images of multiple categories without fine-tuning of parameters;
- We propose a feature-attention module aiming to enhance the differences between normal and abnormal images and increase the focus on significant features.

RELATED WORK

Few-shot learning aims to enable models to recognize classes not seen during the training phase. This requires the model to have a certain degree of generalization ability and an understanding of semantic concepts. Image semantic learning, a key research area in computer vision, is dedicated to enabling computers to not only recognize content in an image but also understand its deeper meaning. Through this combined approach, computers will be able to recognize and interpret visual information more accurately.

Anomaly Detection

The advancement of deep learning technology has significantly enhanced the adoption of deep-learning-based anomaly detection methods. These methods are praised for their robustness in tackling complex situations. Popular techniques include multivariate Gaussian distributions (Dwivedi et al., 2021; Wang et al., 2023), normalizing flow (Rudolph et al., 2021; Papamakarios et al.,2021), k-nearest neighbor algorithms, and Gaussian mixture models (Ran et al., 2021).

The diversity of anomalies underscores the importance of unsupervised anomaly detection. Methods based on embedding similarity (Wang et al. 2023; Zou et al. 2022) compare the distribution modeled by the training set of an image or patch embedding to that of a regular image or patch embedding. These approaches effectively condense regular features into a compact space with anomalous features in the embedding space being distant from the normal clustering. Many methods utilize pre-trained networks from ImageNet for feature extraction. For instance, patch distribution modeling framework for anomaly detection and localization (PaDIM) (Defard et al., 2021) embeds derived anomalous patch features through a multivariate Gaussian distribution using a pre-trained model. PatchCore (Roth et al., 2022) optimizes the representation of nominal patch features using an in-memory repository and assesses input features using either the Mahalanobis distance or the maximum feature distance. However, industrial images often have a different distribution than ImageNet, leading to potential mismatch issues. Additionally, computing the inverse of the covariance or searching for nearest neighbors in a memory bank can negatively impact real-time performance, especially for edge devices.

Image reconstruction-based methods are also widely used in anomaly detection. The assumption here is that anomaly samples cannot be efficiently reconstructed back by compressive reconstruction (Liu et al., 2020) or after mapping to the feature space. Anomalies are identified and classified by comparing the threshold between the reconstruction errors of normal and anomalous samples. Normal image encoding and reconstruction are typically performed for self-encoders and generative adversarial networks (GANs). However, classic autoencoder (AE)-based reconstruction techniques may lose significant features of the original data. Anomaly detection with generative adversarial network (AnoGAN) (Schlegl et al., 2017), one of the pioneering GAN-based methods for image anomaly detection, assesses the resemblance between the discriminator-generated and original data to detect anomalies. However, it is computationally expensive in terms of training time and prone to instability during model training. Recently, new strategies have been developed that involve synthesizing pseudo-defective data from normal samples and training the network in a supervised manner, but there is no guarantee that the trained network will only recreate the normal sample region.

In general, stream-based anomaly detection systems use normalized streams to estimate the density of normal data with non-normal data serving as a low likelihood estimate. Normalized streams can transform any complex distribution into a tractable basis distribution, such as a Gaussian distribution. Early research on normalized streaming models focuses on obtaining feature representations of normal images from pre-trained feature extractors and then learning normally distributed data to focus on local relevance rather than semantic relevance. A current trend is to extract image patches using pre-trained feature encoders. Normalizing flow (NF) (Zavrtanik et al., 2021) is proposed by CS-Flow, anomaly detection with conditional normalizing flow (CFLOW-AD) (Gudovskiy et al., 2022), and DiffNet (Rudolph et al., 2021) to transform normal feature distributions into Gaussian distributions. These approaches are resource-intensive, as normalizing flow can only handle full-size feature mapping without down-sampling, and the coupling layer requires significantly more memory than a regular convolutional layer.

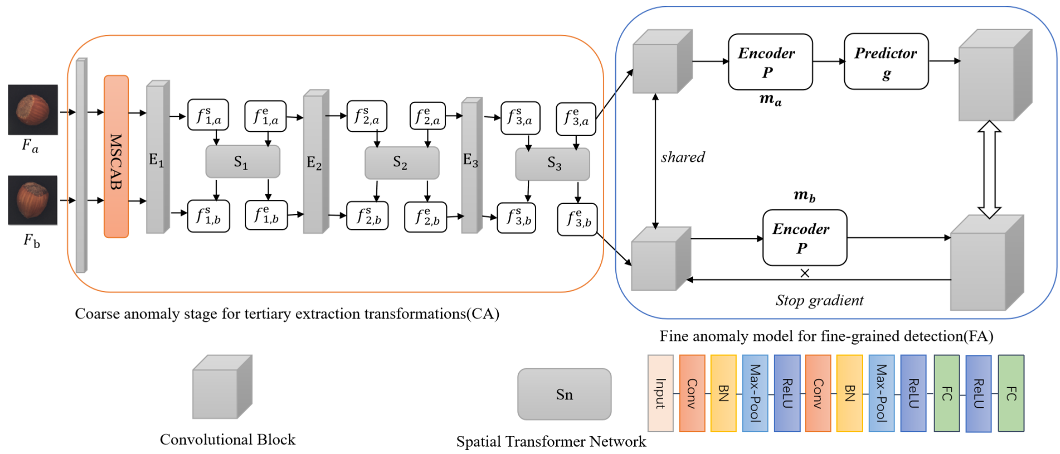
Few-Shot Learning

Unlike traditional anomaly detection methods, which require a large amount of training data for each category, few-shot anomaly detection methods only need a small number of samples for each category. This significantly reduces the cost and time involved. Recently, embedding-based methods have been used to compare embedded vectors in the query set with reference ones. Image reconstruction-based (Fang et al., 2023) methods have shown better results. For example, masked auto-encoder for anomaly-detection(MAEDAY) (Schwartz et al., 2024) recovers missing images in anomaly detection with fewer samples using the self-encoder MAE and achieves good results through pre-training and fine-tuning. Bias networks can be efficiently utilized with a small amount of labeled data for end-to-end anomaly scoring learning on multivariate and image data, respectively.

The development of few-shot anomaly detection is still in its early stages with two main setups: meta-learning and learning that relies on a small number of normal image samples. Meta-learning aims to train a model on various learning tasks so it can solve new learning tasks with only a few training samples (Finn et al., 2017). The training model in this method is easily fine-tuned and often requires a recurrent training model or a Siamese network to be combined with convolutional neural networks (CNNs), full connectivity, or loss functions. On the other hand, methods that rely on a small number of normal image samples use feature learning as the first step in model training, which is then used to compute the anomaly score.

Due to the wide applicability of the registration network, it can be extended across various categories. Therefore, the combination of registration and few-shot anomaly detection is more effective. Based on the above theory, we propose a method to learn the feature distribution.

Figure 2. Anomaly Detection Structure Based on Two-Stage Feature Extraction Transformation



Semantic Learning

Semantic learning aims to enable machines to understand and process semantic information in human language (Hu et al., 2022; Zhou et al., 2022). It has a wide range of applications including industrial inspection (Cvitić et al., 2021; Li & Su, 2022; Mishra,Gupta, et al., 2021; Zemmouchi-Ghomari, 2021), healthcare (Capuano et al., 2022), and educational research (García et al., 2022). Some researchers start from small regions to better establish the association between image features and attributes at a fine-grained level. Dense attribute-based attention mechanisms (Khan et al., 2023) allow the model to better understand and represent unseen categories, while regional attention mechanisms combined with CNNs improve the model's ability to comprehend and process different semantic information in images. Other researchers build regional graphs to capture the relationships between local areas in images, extracting richer semantic information. These methods have the potential to enhance the performance of few-shot learning.

In weakly supervised semantic segmentation (WSSS) methods, boundary exploration approaches also provide positive inspiration for few-shot learning (Li-yi et al., 2020). By improving the model's understanding of the shape and edges of target objects, the visual features of unseen categories can be better understood. Methods based on latent semantic embedding utilize the similarities and semantic relationships between images to enhance classification performance (Luo et al., 2021), allowing for the learning of more robust and effective feature representations with only partially annotated data.

METHOD

In this section, we detail our model framework, depicted in Figure 2. In the third section, we describe the two-stage coarse-to-fine process of few-shot anomaly detection. Afterwards, we explain how the model integrates into a registration network. Then we examine feature distribution through pixel location-based modeling. Finally, we introduce the Gaussian distribution image model resulting from feature extraction and transformation.

Coarse Anomaly (CA) Stage for Tertiary Extraction Transformations

The first stage of coarse anomaly (CA) detection aims to locate all images in a batch of sample images to a consistent spatial position and direction through geometric transformation to obtain images more suitable for feature extraction and subsequent analysis. Unlike traditional methods, semantic granularity is achieved by estimating the rotation angle of the image or by randomly attaching and removing local regions. This technique guarantees an accurate understanding of the image content at each corresponding point. Drawing on the idea that the registration network mainly aligns the input image with the reference image and considering that the training data set is not very large. To avoid the over-fitting problem, we make full use of ResNet to extract the edge and texture features of the image and choose ResNet-18 (Finn et al., 2017) with a shallow network layer as the feature extractor.

Two images of the same category, F_a and F_b, are randomly selected from the training set and input into the feature extractor. Moreover, since the feature maps have different resolutions at the level of the CNN, an interpolation method is required to unify the resolution of the image before the extraction is performed. To ensure that the network possesses spatial transformation capability and efficiently performs fundamental operations like convolutional feature extraction, pooling, and feature compression, we have selected the initial three residual blocks of ResNet-18, expressed as E_1, E_2, and E_3. These blocks are crucial for extracting features at lower layers while maintaining the network's ability to transform spatial information. Meanwhile the image obtained by extracting features is expressed as $f_{1,a}^{s}$ and $f_{1,b}^{s}$. Drawing from the concept of the registration network, a spatial transformation network (STN) is integrated into every residual block (Jaderberg et al., 2015). This integration enables the neural network to dynamically alter the spatial layout of feature maps. Through the application of spatial transformations such as rotation, cropping, and scaling on input data, the model becomes proficient in handling a wide array of geometric variations. This adaptability facilitates enhanced categorization of samples and improved recognition of critical semantic features.

This module does not require supervision, enabling more flexible learning of feature registration, and ultimately, resulting in the generation of the registered image. This strategy skillfully combines the power of deep learning (with ResNet-18) and the flexibility of STN to create an efficient and accurate image alignment and feature matching module. This well-constructed solution not only overcomes the problem of limited training dataset size but also meets the stringent requirements for accurate spatial localization in image analysis, demonstrating its excellent performance in image processing. We regressed the input image to generate an affine transformation function as shown in Equation 1:

$$\begin{pmatrix} x_i^{e} \\ y_i^{e} \end{pmatrix} = \mathrm{T}_\theta\left(f_i^{s}\right) = \mathrm{A}_\theta \begin{pmatrix} x_i^{s} \\ y_i^{s} \\ 1 \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{pmatrix} \begin{pmatrix} x_i^{s} \\ y_i^{s} \\ 1 \end{pmatrix} \tag{1}$$

Among Equation 1, (x_i^{e}, y_i^{e}) is the target coordinate of the output feature $f_{i,m}^{e}$, (x_i^{s}, y_i^{s}) is the source coordinate of the sampling point defined in the input feature mapping f_i^{s}, A_θ is the affine transformation matrix, and T_θ is used to learn the feature mapping of module E. At the same time, a simple model $\mathcal{T}_{\mathrm{T}_\theta}$ is set to learn the content after the above affine transformation, and the ℓ_2 distance of this stage is minimized, as shown in Equation 2:

$$\mathcal{L}_{CA}\left(\mathcal{D};\, \theta_\hbar, \mathrm{T}_\theta\right) = \sum_{f\in\mathcal{D}} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left\| \hbar_{\mathrm{T}_\theta}\left(f_a^{s}\right) - \hbar_{\mathrm{T}_\theta}\left(f_b^{s}\right) \right\|_2 \tag{2}$$

By incorporating an STN, we not only enhance the flexibility of alignment but also bolster the network's ability to adapt to intricate spatial transformations, this integration empowers the network to autonomously learn and adjust the feature map, effectively achieving precise alignment of global embedding positions. This approach allows the network to dynamically adjust its feature map based on the specific requirements of the task at hand.

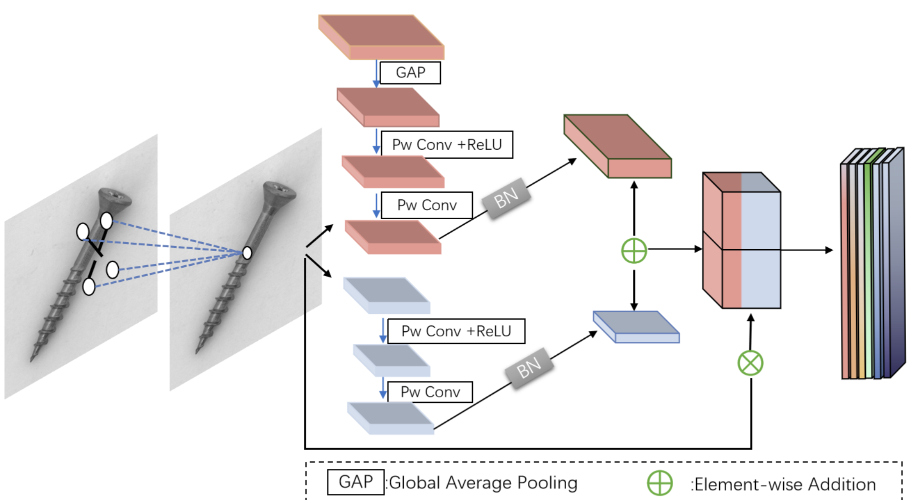
To summarize, the inclusion of an STN in a CNN enables self-learning and autonomous adjustment of the feature map, leading to improved alignment flexibility and enhanced adaptability to complex spatial transformations.

Multiscale Channel Attention Blocks (MSCAB)

To centralize the task-related areas, that is, the most important parts of the network, the anomaly detection and location tasks are enhanced by modeling the importance of each feature channel. Therefore, we have designed a special module to strengthen the focus on features. The multiscale channel attention blocks (MSCAB) proposed by us is composed of two parts, one is composed of a ReLU-controlled masked convolution layer and a multiscale channel attention module (MSCAM), and the other is a channel attention squeeze-and-excitation (SE) module. In the first part, the masked convolution layer is used to perform zero filling of the image after filtering through mask convolution. The masking mechanism guides the network to focus on the key feature regions. Convolution is used at the four sensory field locations and the values are merged and placed in the central location.

The MSCAM module is actually similar to the SE module, which is weighted by the attention of the channel by a single feature to pay more attention to the channel output by the mask convolution. It aggregates multiscale context information along the channel dimension. Finally, the entire module uses a fusion module that implements attention features instead of SE attention blocks to pay more attention to local and global attention. In deep CNNs, there are often correlations and dependencies between different channels. The MSCAB emphasizes both large objects with more global distribution and small objects with more local distribution. In the scenario with fewer samples, MSCAB guides the model to learn valuable and diversified features more accurately and enhance the representation of semantic features. For the detailed structure, please refer to the structure (Dai et al., 2021) in Figure 3. As another part of MSCAB, the SE module automatically determines the importance of each feature channel through learning, which in turn reinforces the features that are beneficial to the task at hand while suppressing those that are less important based on this importance. This mechanism helps to improve the performance of the model on specific tasks and ensures that resources are focused on the most valuable features.

Figure 3. The Detail of the MSCAM Module



In the training phase, since normal images are used, the inputs to the MSCAB are features that have been convolved with the first layer of ResNet. After the pooling operation, activating the mask convolutional layer of ReLU will enhance the attention to the important features and subsequently further strengthen the attention to these key features through the channel attention block. This design helps the network to capture and utilize the key information in the image more accurately, thus enhancing the quality of the feature representation.

Fine Anomaly (FA) Stage for Fine-Grained Detection

Since data enhancement commonly occurs in the process of anomaly detection, data enhancement plays an important role in expanding the dataset, which generally includes translation, rotation, graying, and random combinations of these. As a result, the spatially transformed output image $f_{3,a}^{e}$ and $f_{3,b}^{e}$ from the coarse anomaly detection has to undergo a series of stochastic image enhancement processes before it is fed into the Siamese network. In this network architecture, the top and bottom encoders are designed to be identical to ensure that they provide equivalent feature representations when processing the image. On one side, a predictor is applied, while on the other side, gradient stopping is applied, as shown in Figure 2. The goal is to assign a unique vector representation to each position in the feature map, with the vectors having a smaller distribution range in normal images. Instead of fully connected layers, 1x1 convolutions are used in the feature extractor. The vectors are passed through a shared parameter 3-layer 1x1 convolutional encoder P, and only E($f_{3,a}^{e}$) $\triangleq m_a$ passes through a 2-layer 1x1 predictor g. The output vectors are denoted as m_a=E($f_{3,a}^{e}$) and m_b =E($f_{3,b}^{e}$). Negative cosine similarity is used to represent similarity information for robustness enhancement, as it can perform a “low-budget version of per-pixel registration images” to show the relationship between features of the same dimension in different branches. This is expressed as a feature-level registration loss that is minimized by their negative cosine similarity. After minimizing their negative cosine similarity, it is represented in Equation 3:

$$\mathcal{D}\left(m_a;\ m_b;\ \theta_g,\theta_f\right) = -\frac{g\left(m_a\right), m_b}{\|m_a\|_2 \cdot \|m_b\|_2} \tag{3}$$

Among Equation 3, $\| \ \|_2$ is a l_2 paradigm where g represents the predictor in the Siamese network, θ_g and θ_f are the parameters of the encoder and predictor, respectively. To improve the readability of this formula, we further supplement this equation by expressing it as: $p_a = P\left(E\left(f_{3,a}^{e}\right)\right) = P\left(m_a\right)$, $m_b = E\left(f_{3,b}^{e}\right)$. Finally, a symmetry of the feature registration loss is defined, as formulated in Equation 4:

$$\mathcal{L} = \frac{1}{2}\mathcal{D}\left(p_a, m_b\right) + \frac{1}{2}\mathcal{D}\left(p_b, m_a\right) \tag{4}$$

An important component of the method is the stop gradient operation, which changes the previous formula as shown in Equation 5:

$$\mathcal{L} = \frac{1}{2}\mathcal{D}\left(p_a, stopgradient\left(m_b\right)\right) + \frac{1}{2}\mathcal{D}\left(p_b, stopgradient\left(m_a\right)\right) \tag{5}$$

In Equation 5, m_b and m_a become two constants. The stop-gradient operation plays a crucial role in preventing the gradient vanishing problem, which is key to avoiding numerically unstable solutions. This technique effectively halts the gradient flow, thereby maintaining the robustness and reliability of the network's learning process. It acts as a safeguard within the training regimen, ensuring that the evolution of the model's parameters is controlled and predictable, ultimately leading to more stable and reliable outcomes.

Feature Distribution

At each pixel position of the feature map, we use the Gaussian distribution to model the representation of the normal distribution of the target category features. By modeling the feature distribution, a probability distribution can be provided for each pixel in the feature map. Since the two branches of the twin network are the same, only one branch feature is used to estimate the normal distribution, assuming the image feature distribution position after the divided grid location is set to $\left(i, j\right)$. Firstly, the patch embedding vector at the position $\left(i, j\right)$ is calculated. Then, X_{ij} becomes from N random enhanced support image set extraction features, where $\mathrm{X}_{ij} = \left\{f_{ij}^k, k \in \left[1, N\right]\right\}$. In order to summarize the information carried by the proposed model, we construct a multivariate Gaussian distribution $\mathcal{N}\left(\mu_{ij}, \Sigma_{ij}\right)$, where the μ_{ij} is the mean sample of X_{ij} . X_{ij} is defined as the aggregated feature of the patch position, and the sample covariance is defined in Equation 6:

$$\Sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}\left(f_{ij}^k - \mu_{ij}\right)\left(f_{ij}^k - \mu_{ij}\right)^T + \xi I \tag{6}$$

The covariance matrix $£_{\mathrm{ij}}$ computation is rooted in the concept of sample covariance, a measure that captures the spread and inter-dependence of features. This process begins by determining the deviation of each eigenvector f_{ij}^k from its mean vector μ_{ij}. These deviations are then subjected to a dot-product operation, which quantifies the similarity in direction between the vectors. The results of these dot-products are accumulated and normalized by the number of samples minus one, yielding an unbiased estimate of the covariance. The resulting covariance matrix $£_{\mathrm{ij}}$ provides a comprehensive depiction of the sample points' distribution within the feature space, encapsulating both the variance along each dimension and the inter-dimensional correlations. This matrix is a pivotal descriptor in understanding the underlying structure of the data and is instrumental in various machine learning tasks, such as classification and clustering.

Inference

After intensive feature extraction, we model the normal image with Gaussian distribution for each image during the test process. For the test image, we extract the normal image representation of the position $\left(i, j\right)$, and calculate the anomaly score using Mahalanobis distance for the patch, as shown in Equation 7:

$$A_{ij} = \sqrt{\left(x_{ij} - \mu_{ij}\right)^{\mathrm{T}} \Sigma_{ij}^{-1}\left(x_{ij} - \mu_{ij}\right)} \tag{7}$$

Among Equation 7, assuming that the image is divided into grid positions $\left(i,j\right)$, and the normal distribution feature resolution is $\mathrm{W}\times\mathrm{H}$, we have established a special Gaussian distribution for each position $\left(i,j\right)_{1\leq i\leq W,1\leq j\leq H}$ to form an anomaly map in Equations 8 and 9:

$$\mu_{ij}=\frac{1}{N}\sum_{k}^{N}x_{ij}^{k} \tag{8}$$

$$\Sigma_{ij}=\frac{1}{N-1}\sum_{k}^{N}\left(x_{ij}^{k}-\mu_{ij}\right)\left(x_{ij}^{k}-\mu_{ij}\right)^{\mathrm{T}} \tag{9}$$

EXPERIMENTAL RESULTS AND ANALYSIS

Datasets

In this study, we employ the MVTec Anomaly Detection (MVTec AD) dataset (Bergmann et al., 2019) to validate the effectiveness of our proposed model, and then we assess the model's generalization performance on the bearing test anomaly detection (BTAD) dataset (Mishra, Verk, et al., 2021).

MVTec AD is a publicly available dataset widely used for industrial anomaly detection research. This dataset provides high-quality images of various industrial products for training and testing anomaly detection algorithms. Anomalies in the dataset typically occur during the manufacturing process and are critical for the development of automated visual inspection systems. The diversity and complexity of the MVTec AD dataset make it an ideal choice for evaluating and comparing the performance of anomaly detection algorithms. The MVTec AD contains 15 categories of industrial products, including 10 types of physical objects and five types of textures. Each category comprises a varying number of images, totaling 5,354 images. There are 3,629 defect-free images for training and 1,725 images for testing, which include both defect-free samples and samples with various types of defects, such as scratches, dents, and stains. The image resolution in the MVTec AD dataset ranges from 700×700 to 1024×1024 pixels.

The BTAD dataset is a real-world industrial anomaly detection dataset, comprising 2,830 authentic images of three different industrial products. The images have resolutions of 1600×1600, 600×600, and 800×600 pixels, respectively. It provides a comprehensive set of features extracted from vibration signals, which are used for training and evaluating anomaly detection models. The diverse range of fault conditions within the BTAD dataset makes it an ideal choice for assessing the generalization capability of such models in detecting unseen anomalies in bearing systems.

Implementation Details

We implement the model in Pytorch. The model trains 50 epochs on a single Nvidia 3090 with a batch size of 16. We update the parameters using momentum SGD with a learning rate of 0.001. Meanwhile we set the training weight attenuation to $5e^{-2}$, and the training time of a class takes about eight hours.

In this study, we employ the area under the receiver operating characteristic curve (AUCROC) as the benchmark for evaluating the performance of image anomaly detection and pixel-level anomaly localization (Xie et al., 2024). The AUROC metric ranges from 0 to 1.

In the preprocessing stage, we resize the images of all categories to 224×224. The training set is composed of n normal images of known categories, that is, $\mathcal{T}_{train}=U_{i=1}^{n}\mathcal{M}_{i}\left(i=1,\ldots n\right)$;

Table 1. Comparison With Other Methods at the Image-Level in AUCROC (%)

Category	TDG	k=2		ours	TDG	k=4		ours	TDG	k=8		ours
		DiffNet+	RegAD			DiffNet+	RegAD			DiffNet+	RegAD	
Bottle	69.3	99.3	99.4	**99.5**	69.6	99.3	**99.4**	**99.4**	70.3	99.4	99.8	**99.9**
Cable	68.3	**85.3**	65.1	72.2	70.3	**85.2**	76.1	80.7	74.7	**87.9**	80.6	85.2
Capsule	55.1	**73.0**	67.5	70.7	47.6	**80.3**	72.4	75.9	44.7	**78.6**	76.3	77.9
Carpet	66.2	78.4	**96.5**	93.0	68.7	78.6	**97.9**	95.0	78.2	78.5	**98.5**	96.6
Grid	83.8	62.1	84.0	**84.3**	86.2	60.5	91.2	**91.5**	87.6	78.5	91.5	**91.6**
Hazelnut	67.2	94.9	96.0	**98.4**	71.2	95.8	95.8	**98.2**	82.8	97.9	96.5	**98.4**
Leather	93.6	90.7	99.4	**99.5**	93.2	91.2	**100.0**	**100.0**	93.5	92.2	**100.0**	**100.0**
Metal Nut	67.1	61.9	**91.4**	82.9	69.2	67.3	**94.6**	92.5	68.7	67.6	**98.3**	94.2
Pill	69.2	83.2	**81.3**	77.8	64.7	**84.0**	80.8	81.7	67.9	82.1	80.6	**82.5**
Screw	**98.8**	73.4	52.5	79.0	**98.8**	72.5	56.6	77.6	**99.0**	75.0	63.4	78.4
Tile	86.3	97.0	94.3	**97.4**	87.2	**98.0**	95.5	96.6	87.4	**99.6**	97.4	98.6
Toothbrush	54.4	60.8	**86.6**	83.2	57.8	62.5	90.9	**96.1**	57.6	60.8	98.5	**99.1**
Transistor	55.9	61.8	**86.0**	80.5	67.7	62.2	85.2	**88.4**	71.5	63.3	**93.4**	89.2
Wood	98.4	98.1	99.2	**99.7**	98.3	96.4	98.6	**98.8**	98.4	99.4	99.4	**99.7**
Zipper	64.4	**89.2**	86.3	81.7	65.3	84.8	88.5	**88.6**	66.3	87.3	**94.0**	92.8
Average	73.2	80.6	85.7	**86.7**	74.4	81.3	88.2	**90.7**	76.6	83.2	91.2	**92.3**

Note: k denotes the number of shots in our few-shot settings.

its subset $\mathcal{M}$ is composed of normal images in the category $\mathcal{C}_i\left(i=1,2,\ldots n\right)$. During testing, a few normal samples of the target category $\mathcal{C}_t\left(t \notin \left(i=1,2,\ldots n\right)\right)$ are available and supported by a few normal images.

Comparisons With Other Methods

In this experiment, one target category is used for testing, and the other remaining categories are used for training. Table 1 shows the comparison of our method with other methods on the MVTec AD dataset at k=2, k=4, and k-8, respectively. In terms of image-level anomalies, we compare our method with TDG (Sheynin et al., 2021), DiffNet (Rudolph et al., 2021), and RegAD (Huang et al., 2022) methods, all of which use the same ImageNet pretraining architecture. The advantage of our proposed method is that it improves the adaptability of the model to various transformations by using feature alignment in the image registration module and shows superior performance in most categories.

Moreover, it is easy to find that our method achieves the best average AUROC in all three scenarios. In k=2 and k=8, our method outperforms other methods in six out of 15 categories. At the same time, in terms of image-level, average results, our method improves accuracy by 1% compared to the state-of-the-art (SOTA) method. At k=4, our method outperforms other methods in eight categories, and the average results at the image level improve by 2% compared to SOTA. This is enough to prove the effectiveness of our method.

Compared with the SOTA model, our method improves the average area under the curve (AUC) in the MVTec dataset by about 1%, 2.5%, and 1.2% in the cases of k=2, k=4, and k=8, respectively. It shows good performance in bottles, hazelnuts, leather, wood, etc. Specifically, for texture-rich classes such as wood and leather, color enhancement is deemed inappropriate for these images; hence, we abstain from applying such enhancements, which leads to commendably robust, image-level outcomes. Meanwhile, it has sub-optimal accuracy in cables, capsules, and metal nuts. Furthermore, certain categories are not amenable to specific feature-level spatial manipulations, as seen with transistors where operations like flips or rotations do not yield favorable results. The data in the table clearly indicate that the performance of transistor images following affine transformations is suboptimal compared to other established methods. Due to the model's use of two-stage, image-feature-level and pixel-level features for learning and alignment, the concept of feature alignment can be generalized to different categories for general feature learning, improving the model's generalization ability. It can be easily observed from Figure 4 that, from 1-shot to 8-shot, our method outperforms the registration based few-shot anomaly detection (RegAD) method, as well as the PaDIM (Defard et al., 2021) and student-teacher feature pyramid matching (STPM) (Wang et al., 2021) methods in image-level results. Among them, k represents using several corresponding category images as the training set, rather than a few images. On this dataset, the gap between us and the baseline is also the largest, indicating that our model utilizes the additional variability of the samples. As the number of class images increases, the performance of the model also improves. Therefore, when we use a small sample setting for anomaly detection, our method outperforms other methods in image-level results.

Considering that other methods use a few images as training sets rather than several categories, our approach adapts the model to unseen categories after training on multiple categories. We use a few samples as the support set. Considering fairness, we reimplement the SOTA method with the number of k, such as semantic pyramid anomaly detection (SPADE) (Cohen & Hoshen, 2020), STPM (Wang et al., 2021), RD4AD (Deng & Li, 2022), and coupled-hypersphere-based feature adaptation (CFA) (Lee et al., 2022), using the official source code for the experiments. Meanwhile, the sampling rates are all set at 0.001. The comparison results of the MVTec AD dataset in terms of image level and pixel level are also shown in Table 2.

In the experiment, we compare our approach with some classical anomaly detection methods. Researchers typically use pixel-level metrics to evaluate the performance of anomaly detection methods. Fundamentally, pixel-level metrics focus on the ability to identify anomalies within an image, while image-level metrics focus on determining whether a product, as a whole, is anomalous. In practice, the lower the defect level, the higher the price of the product. In addition, as shown in Table 2, some few-shot anomaly detection methods, like CFA, perform well on pixel-level metrics but poorly on image-level AUROC results. In our opinion, both metrics demonstrate different capabilities for anomaly detection that can greatly benefit industrial manufacturing production.

Therefore, it is important to develop an unsupervised anomaly detection method that excels in both pixel-level and image-level results without the need for labeled samples.

Compared to the other methods in Table2, our method outperforms the other baseline methods in the MVTec AD dataset in both image-level and pixel-level average AUROC. At the image level at k = 2, our method outperforms the other methods in 10 out of 15 categories, and again at the pixel

Figure 4. Comparison With Other Methods at the Image-Level in AUCROC (%)

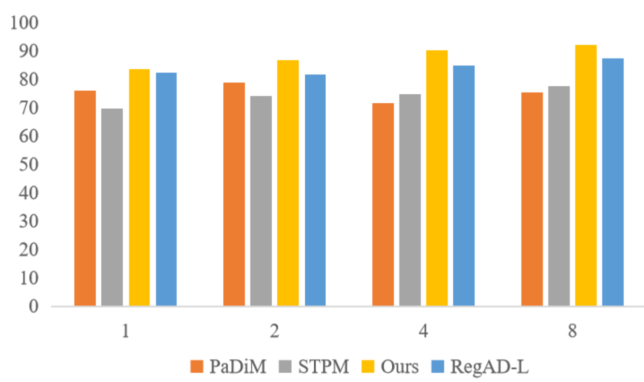


Table 2. Performance Comparison With SOTA Methods for Anomaly Detection and Localization

	Category	CFA	SPADE	STPM	RD4AD	Ours
k=2	Bottle	96.7/93.2	95.2/85.2	93.2/84.3	91.2/81.7	**99.5/95.5**
	Cable	65.4/88.2	60.1/78.2	59.8/50.9	58.3/64.8	**72.2/93.6**
	Capsule	50.2/85.6	45.6/79.2	43.2/49.2	44.7/77.9	**70.7/70.6**
	Carpet	**97.1**/97.5	93.2/95.2	90.5/60.5	92.5/72.5	93.0/**97.6**
	Grid	79.2/**81.3**	75.1/75.6	71.2/61.2	74.3/74.3	**84.3**/51.8
	Hazelnut	98.1/**98.1**	95.0/88.2	90.3/73.3	93.2/63.2	**98.4**/98.0
	Leather	**100/99.2**	97.2/88.3	95.1/75.1	96.5/86.5	99.5/99.0
	Metal_nut	66.1/89.5	60.2/58.5	58.2/51.1	63.4/68.7	**82.9/96.9**
	Pill	66.3/91.2	59.7/54.2	57.3/49.3	62.4/65.6	**77.8/96.3**
	Screw	55.9/**96.5**	49.6/69.6	51.2/51.2	53.5/59.7	**79.0**/95.8
	Tile	**99.8**/81.5	89.5/81.5	90.2/57.2	88.7/88.7	97.4/**89.4**
	Toothbrush	**86.7**/93.8	78.5/75.5	75.2/65.2	77.8/77.8	83.2/**94.7**
	Transistor	71.5/79.5	50.5/73.5	**83.2**/43.2	77.5/77.5	80.5/**94.3**
	Wood	98.1/91.8	49.5/89.5	95.4/45.4	93.5/**93.5**	**99.7**/93.4
	Zipper	50.3/93.2	48.5/**93.5**	45.2/55.2	48.6/48.6	**81.7**/91.8
	Average	78.8/**90.6**	69.8/79.1	69.7/58.2	74.4/69.0	**86.7/90.6**
k=4	Bottle	94.2/93.6	95.8/86.9	93.9/84.9	92.1/81.8	**99.4/97.9**
	Cable	**91.2**/89.1	61.3/78.7	61.3/52.2	68.4/66.2	80.7/**95.9**
	Capsule	56.2/86.2	48.7/80.1	47.4/59.3	51.7/78.4	**75.9/98.1**
	Carpet	**97.6/98.2**	92.5/95.0	91.5/60.6	93.2/74.8	95.0/97.9
	Grid	81.5/**82.5**	76.2/76.1	75.3/61.8	76.4/76.9	**91.5**/66.3
	Hazelnut	**99.4**/98.5	95.6/89.1	91.4/74.9	93.8/65.2	98.2/**98.8**
	Leather	**100.0/99.3**	98.2/89.3	96.9/75.3	96.8/86.7	**100.0**/99.2
	Metal_nut	91.3/89.9	62.5/60.2	60.8/51.8	65.3/69.2	**92.5/97.9**
	Pill	**85.6**/91.6	61.8/58.2	61.3/50.6	62.8/70.4	81.7/**97.2**
	Screw	49.2/**96.8**	52.9/71.3	52.8/51.9	55.7/60.9	**77.6**/95.3
	Tile	**99.8**/82.3	91.3/82.4	90.4/58.5	90.8/59.5	96.6/**94.1**
	Toothbrush	87.2/94.2	81.7/76.9	80.4/66.9	76.7/78.9	**96.1/97.2**
	Transistor	**95.8**/80.5	52.5/74.2	82.4/57.5	79.3/67.9	88.4/**93.2**
	Wood	98.6/92.6	51.4/90.4	95.8/48.9	94.2/**94.2**	**98.8/94.2**
	Zipper	**94.3**/94.8	52.2/93.8	47.6/56.4	56.7/52.3	88.6/**97.0**
	Average	85.0/91.3	71.6/80.2	74.77/60.8	76.9/72.2	**90.7/94.7**

continued on following page

Table 2. Continued

	Category	CFA	SPADE	STPM	RD4AD	Ours
k=8	Bottle	95.1/93.6	95.9/87.1	94.1/85.2	92.8/82.1	**99.9/98.4**
	Cable	**91.8**/89.2	63.5/78.9	62.6/53.3	69.2/68.2	85.2/**96.6**
	Capsule	69.5/86.5	58.9/80.2	57.8/59.3	58.5/78.5	**77.9/98.2**
	Carpet	**97.6**/98.4	92.7/95.1	91.6/60.7	93.8/79.2	96.6/**98.3**
	Grid	85.6/**82.8**	77.3/77.2	76.9/61.8	77.9/76.9	**91.6**/67.6
	Hazelnut	99.4/98.6	96.5/89.5	91.8/74.9	94.2/65.5	**98.4/98.8**
	Leather	**100.0**/99.4	98.7/90.2	97.2/75.3	97.2/86.9	**100.0/99.9**
	Metal_nut	92.3/89.9	68.9/60.5	61.3/54.6	65.6/69.5	**94.2/98.2**
	Pill	**88.9**/91.7	63.9/58.2	64.2/55.7	63.6/70.5	82.5/**98.1**
	Screw	65.4/**96.9**	56.4/71.4	55.9/52.3	59.3/61.9	**78.4**/95.5
	Tile	**99.8**/83.4	91.8/82.5	91.2/58.9	91.2/60.8	98.6/**91.8**
	Toothbrush	88.9/94.5	82.9/77.2	82.3/66.9	77.9/79.1	**99.1/97.9**
	Transistor	**96.2**/81.5	58.9/74.5	84.6/58.2	81.2/67.9	89.2/**97.0**
	Wood	98.9/92.7	61.3/90.4	95.8/49.2	95.6/94.5	**99.7/95.4**
	Zipper	**94.5**/94.9	62.5/94.2	57.2/57.8	58.9/52.8	92.8/**97.7**
	Average	90.9/91.6	75.3/80.5	77.6/61.6	78.5/73.0	**92.3/95.3**

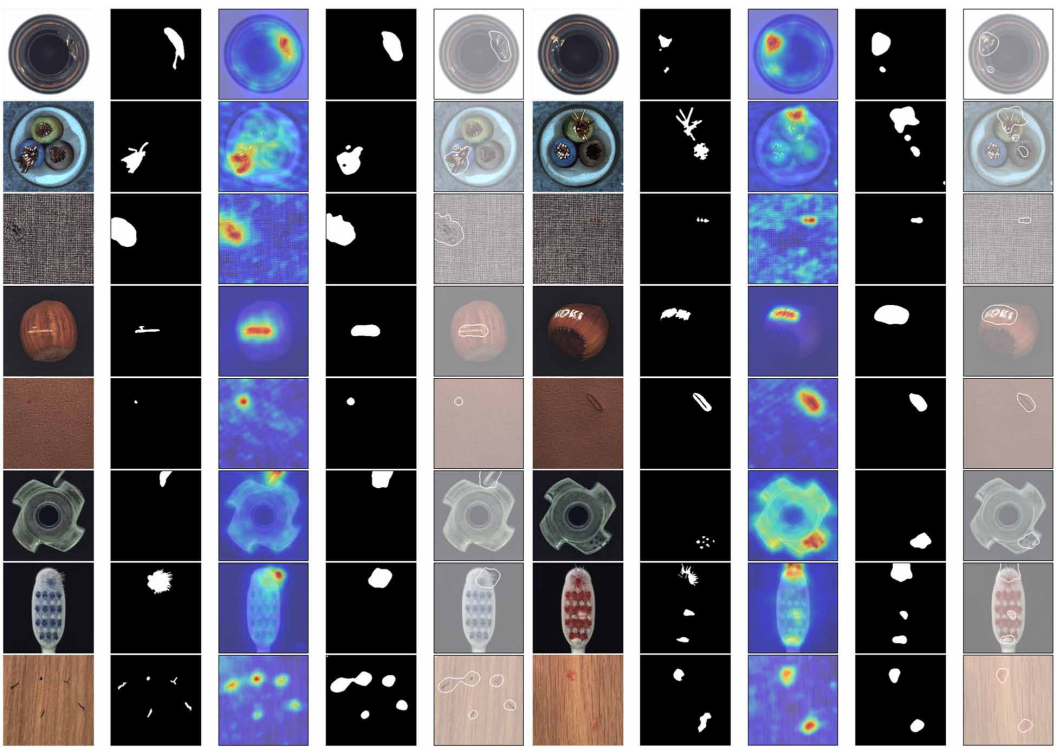
level, our method performs better in nine categories. The method in this study is trained and tested without involving parameter fine-tuning, a situation that may not guarantee optimal performance for each category. Other baseline methods can adjust parameters based on the training situation of each category, thus other baseline methods may have the advantage of allowing the best performance for each category after parameter fine-tuning. In the cases of k=4 and k=8, compared to the four methods in Table 2, our method performs best or is suboptimal at the image and pixel levels. At k=8, our method achieves an image-level AUC of 92.3 during training, which is approximately 2% higher than the CFA method that adapts to dataset features for complex anomaly detection and localization in a small sample setting. Pixel-level metrics are comparative results of anomalous localization. The first number in Table 2 indicates the image-level AUROC, the latter, the pixel-level AUROC. It shows from both tables that our method achieves the best level in the few-sample setting in the average image-level and pixel-level AUROC, for all categories, which is sufficient to demonstrate the effectiveness of our method.

The statistical results from Table 2 and Table 3 imply that our proposed few-shot anomaly detection method from coarse to fine stages has achieved good results in the current work. It has made significant improvements in the defects of objects. Moreover, it can be found that Table 3 clearly lists the comparison between our method and other methods in terms of training data selection, whether it is pre-trained or not, and the selection of the backbone network. Figure 5 illustrates the distribution of anomaly scores for the test sample. The first column represents the original image in the dataset; the second column represents the ground truth map; the third column is the predicted heat map; the fourth column is the predicted mask map; and the last column represents the final segmented result of the image, after processing. Meanwhile, it can also be found that in Figure 5 the anomaly heat map not only highlights the location and shape of the defects, but also observes that the map pays attention to the fine granularity of the image, which shows how our method pays attention to the anomalies.

Table 3. Comparison With Existing Methods in Different Aspects

Methods	Data	ImageNet Pretrain	Backbone	MVTec AD	
				image	pixel
Ours(k=2)	2 images	√	ResNet-18	86.7	90.6
Ours(k=4)	4 images	√	ResNet-18	90.7	94.7
Ours(k=8)	8 images	√	ResNet-18	92.3	95.3
GANomaly	Full images	×	UNet	80.5	-
ARNet	Full images	×	UNet	83.9	-
MKD	Full images	√	ResNet-18	87.7	90.7
CutPaste	Full images	√	ResNet-18	95.2	96
PaDiM	Full images	√	WRN50	97.9	97.5
PatchCore	Full images	√	WRN50	99.1	98.1
CflowAD	Full images	√	WRN50	98.3	98.6
SPADE	Full images	√	WRN50	78.9	90.5

Figure 5. The Distribution of Anomaly Scores When the Test Sample is Plotted



To validate the generalization capability of our proposed model, we conducted experiments on industrial dataset BTAD. In terms of image-level anomalies, we compared our method with PatchCore (Roth et al., 2022), discriminatively trained reconstruction anomaly embedding model (DRAEM) (Zavrtanik et al., 2021), patch-level support vector data description (PatchSVDD) (Yi & Yoon, 2021),

Table 4. Comparison of Our Method With Other Methods on the BTAD Dataset

Method	PatchCore	DRAEM	PatchSVDD	MKD	Our
Image-level	0.93	0.92	0.92	**0.94**	**0.94**
Pixel-level	**0.98**	0.94	0.96	0.97	0.97

and multiresolution knowledge distillation (MKD) (Salehi et al., 2021) methods, and experimental results show the effectiveness of the proposed method. As shown in Table 4, the accuracy of the model at the image level and pixel level reaches 94.3% and 97.2%, respectively. Figure 6 illustrates the test results, highlighting the model's performance in distinguishing between normal and anomalous instances within the dataset.

Ablation Study

The ablation experiments described below are all performed on the MVTec AD dataset, and some other parameters were consistent. We explore the role of important parts of our approach. "Baseline" refers to the use of only the ImageNet, pre-trained ResNet-18, and Augmentation (AUS) stands for the augmentation of the support set samples. The CA stage is the first stage of our proposed two-stage anomaly detection method, which is regularizing the initial image-level features in the sample images; the fine anomaly (FA) stage maximizes the feature similarity of the corresponding position on each image. MSCAB enhances the difference between normal images and samples by adding global and local channel attention. We verified the influence of STN and MSCAB on our experiment in Table 5. From Table 5, it is evident that incorporating enhancement processing into the samples using the

Figure 6. The Anomaly Localization Results on BTAD Dataset

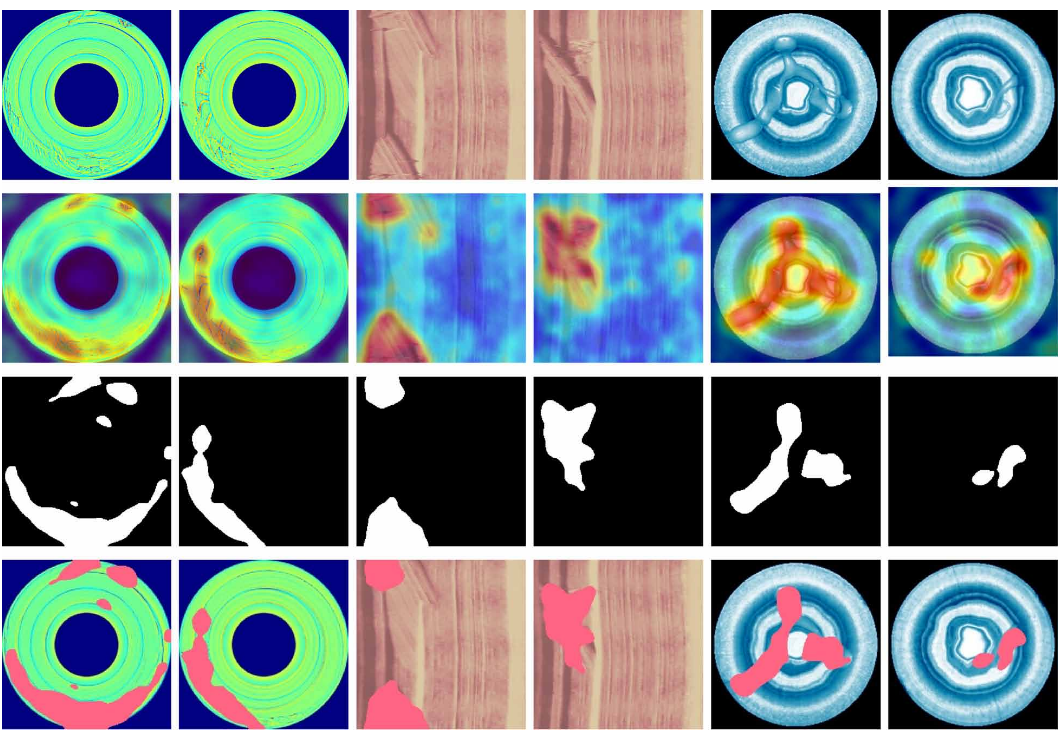


Table 5. Results of Ablation Experiments on the MVTec AD Dataset

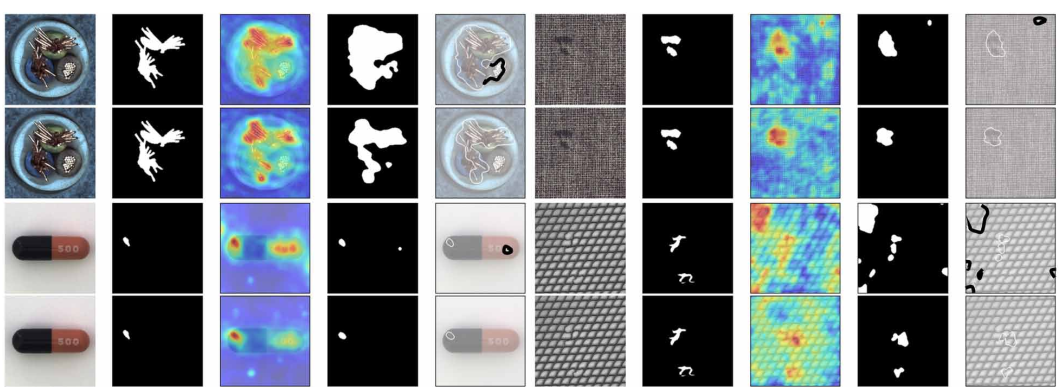
Baseline	AUS (Augmentation)	STN	MSCAB	image			pixel		
				k=2	k=4	k=8	k= 2	k=4	k=8
✓				74.7	78.0	80.5	88.6	90.5	92.1
✓	✓			83.0	86.4	89.3	94.7	95.9	96.6
✓	✓		✓	84.3	87.1	90.5	94.0	95.6	95.9
✓	✓	✓		85.7	88.2	91.2	94.6	95.8	96.1
✓	✓	✓	✓	86.7	90.7	92.3	90.6	94.7	95.3

baseline method significantly improves the image-level results and pixel-level results at the three k-values by approximately 8% to 9% and 5% to 6%, respectively. In order to compare the effects of the two stages of STN and MSCAB for feature mapping, this study includes experiments on the baseline plus the enhancement processing method for each stage. The results show both stages have enhancement effects on anomaly detection and localization. The performance of anomaly detection improves when the feature mapping of both stages is increased simultaneously, resulting in an average enhancement of 11%-12% compared to the baseline method. This underscores the significance of enhanced detail region detection for anomaly detection results and localization.

Adding one module separately has an impact on both baseline methods. According to Table 5, compared to the baseline method, adding the STN module alone increases the average image-level AUROC by 10.6% and the average pixel-level AUROC by 5.1%. Meanwhile, compared to the baseline method, adding the MSCAB module alone results in an average increase of 9.5% in image-level results and 4.9% in pixel-level results. The enhancement of model detection and localization capabilities can be intuitively observed in the data.

The results show that both of them have a positive effect on improving the performance of anomaly detection and improving the image-level AUROC. The proposed STN module improves the feature registration capability. The combination of the two can produce better positive feedback results on the image-level results. Figure 7 clearly demonstrates that the baseline approach considers both normal and abnormal features, while our method prioritizes abnormal features. The illustration highlights the contrast in the dataset's four categories when employing the baseline with the enhancement method versus the approach adopted in our study. The bolded box lines in the segmentation result shown in the figure represent the baseline plus enhancement method's attention to the non-anomalous features.

Figure 7. Qualitative Results of Anomaly Localization on the MVTec Dataset



In contrast, our method more accurately locates the abnormal features, and the abnormal areas in the heat maps are more concentrated at the abnormal points. Meanwhile, our approach does not care about the size and number of exception regions; this fully demonstrates the superiority of our method.

CONCLUSION

This article proposes a new two-stage, few-shot anomaly detection network, meticulously designed to enhance the accuracy and robustness of anomaly detection. This method employs a coarse-grained anomaly detection module for feature extraction, a MSCAM for feature transformation, and a fine-grained detection module to intensify the focus on fine-grained features. It addresses the previous challenges faced by CNNs, such as a high sensitivity to noise and a lack of sensitivity to fine-grained features. Experimental results demonstrate this approach achieves an image-level AUROC of 92.3% and a pixel-level AUROC of 95.3% on the MVTec AD dataset. Future research will focus on further optimizing the model structure and algorithms. For example, considering diffusion models to effectively improve the accuracy of low-parameter feature extraction.

In our future research, we will prioritize methods that aim to simplify model structures, thus reducing computational complexity and training costs, while also enhancing interpretability and deployability for industrial anomaly detection applications. Additionally, we plan to incorporate diffusion models into our future research efforts to enhance anomaly detection methods. These models are proficient at capturing spatiotemporal semantic correlations in data, thereby improving the accuracy and robustness of anomaly detection. This initiative will provide more reliable anomaly detection solutions for industrial intelligence and automation, ultimately reinforcing the reliability and efficiency of industrial production processes.

ACKNOWLEDGMENT

This research was supported by the Research Project of Hubei Provincial Department of Education (No. Q20211802).

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

CORRESPONDING AUTHOR

Correspondence should be addressed to Chengzhi Lyu (China, lyuchengzhi@163.com)

FUNDING

No funding was received for this work.

PROCESS DATES

Received: 3/5/2024, Revision: 4/1/2024, Accepted: 4/1/2024

REFERENCES

Atabay, H., & Hassanpour, H. (2023). Semi-Supervised anomaly detection in electronic-exam proctoring based on skeleton similarity. *2023 11th European Workshop on Visual Information Processing*. doi:10.1109/EUVIP58404.2023.10323052

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD — A comprehensive real-world dataset for unsupervised anomaly detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9584–9592. doi:10.1109/CVPR.2019.00982

Bozorgtabar, B., & Mahapatra, D. (2023). Attention-Conditioned augmentations for self-supervised anomaly detection and localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), 14720–14728. doi:10.1609/aaai.v37i12.26720

Capuano, N., Foggia, P., Greco, L., & Ritrovato, P. (2022). A semantic framework supporting multilayer networks analysis for rare diseases. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–22. doi:10.4018/IJSWIS.297141

Cvitić, I., Perakovic, D., Gupta, B. B., & Choo, K.-K. R. (2021). Boosting-based DDoS detection in internet of things systems. *IEEE Internet of Things Journal*, *9*(3), 2109–2123. doi:10.1109/JIOT.2021.3090909

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., & Barnard, K. (2021). Attentional feature fusion. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3559–3568. doi:10.1109/WACV48630.2021.00360

Defard, T., Setkov, A., Loesch, A., & Audigier, R. (2021). *PaDiM: A patch distribution modeling framework for anomaly detection and localization. n: Del bimbo, A., et al. pattern recognition. ICPR international workshops and challenges. ICPR 2021* (Vol. 12664). Lecture Notes in Computer Science, doi:10.1007/978-3-030-68799-1_35

Deng, H., & Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* New Orleans, LA, USA (2022), 9727–9736. doi:10.1109/CVPR52688.2022.00951

Ding, C., Pang, G., & Shen, C. (2022). Catching both gray and black swans: Open-set supervised anomaly detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* New Orleans, LA, USA (2022), 7378–7388. doi:10.1109/CVPR52688.2022.00724

Dwivedi, R. K., Kumar, R., & Buyya, R. (2021). Gaussian distribution-based machine learning scheme for anomaly detection in healthcare sensor cloud. *International Journal of Cloud Applications and Computing*, *11*(1), 52–72. doi:10.4018/IJCAC.2021010103

Fang, Z., Wang, X., Li, H., Liu, J., Hu, Q., & Xiao, J. (2023). FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17435–17444. doi:10.1109/ICCV51070.2023.01603

Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning. Proceedings of Machine Learning Research*, 1126–1135. doi:10.48550/arXiv.1703.03400

García, A. M. F., García, M. I. M., Gallinas, R. B., Sánchez, M. M., & Gutiérrez, M. E. B. (2022). Integration and open access system based on semantic technologies. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–19. doi:10.4018/IJSWIS.309422

Gudovskiy, D., Ishizaka, S., & Kozuka, K. (2022). CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 98–107. doi:10.1109/WACV51458.2022.00188

Hu, B., Gaurav, A., Choi, C., & Almomani, A. (2022). Evaluation and comparative analysis of semantic web-based strategies for enhancing educational system development. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–14. doi:10.4018/IJSWIS.302895

Ilyas, Q. M., Ahmad, M., Rauf, S., & Irfan, D. (2022). RDF query path optimization using hybrid genetic algorithms. *International Journal of Cloud Applications and Computing*, *12*(1), 1–16. doi:10.4018/IJCAC.2022010101

Jaderberg, M., Simony, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. *Advances in Neural Information Processing Systems*, *28*. Advance online publication. doi:10.48550/arXiv.1506.02025

Khan, W., Ullah, L., Khan Bangash, A., Khan, M., & Khan, Y. (2023). Challenges and Potential Opportunities of Tourism in Kumrat Valley, Khyber Pakhtunkhwa. *Qlantic Journal of Social Sciences*, *4*(4), 108–120. doi:10.55737/qjss.319897332

Lee, S., Lee, S., & Song, B. C. (2022). CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 78446–78454. doi:10.1109/ACCESS.2022.3193699

Li, H., Zheng, W., Tang, F., Zhu, Y., & Huang, J. (2023). Few-shot time-series anomaly detection with unsupervised domain adaptation. *Information Sciences*, *649*, 119610–119610. doi:10.1016/j.ins.2023.119610

Li, J., & Su, J. (2022). Semantic trajectory frequent pattern mining model. *International Journal on Semantic Web and Information Systems*, *18*(1), 1–20. doi:10.4018/IJSWIS.313190

Li-yi, C., Wu, W., Fu, C., Han, X., & Zhang, Y. (2020). Weakly supervised semantic segmentation with boundary exploration. *Lecture Notes in Computer Science*, *12371*, 347–362. doi:10.1007/978-3-030-58574-7_21

Liu, R. W., Guo, Y., Lu, Y., Chui, K. T., & Gupta, B. B. (2023). Deep network-enabled haze visibility enhancement for visual iot-driven intelligent transportation systems. *IEEE Transactions on Industrial Informatics*, *19*(2), 1581–1591. doi:10.1109/TII.2022.3170594

Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R. J., & Camps, O. (2020). Towards visually explaining variational autoencoders. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR42600.2020.00867

Luo, W., Yang, M., & Zheng, W. (2021). Weakly-supervised semantic segmentation with saliency and incremental supervision updating. *Pattern Recognition*, *115*, 107858. doi:10.1016/j.patcog.2021.107858

Mishra, A., Gupta, N., & Gupta, B. B. (2021). Defense mechanisms against DDoS attack based on entropy in SDN-cloud using POX controller. *Telecommunication Systems*, *77*(1), 47–62. doi:10.1007/s11235-020-00747-w

Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., & Foresti, G. L. (2021). VT-ADL: A vision transformer network for image anomaly detection and localization. *2021 IEEE 30th International Symposium on Industrial Electronics*, 01–06. doi:10.1109/ISIE45552.2021.9576231

Pang, G., Yan, C., Shen, C., Hengel, A., & Bai, X. (2020). Self-Trained deep ordinal regression for end-to-end video anomaly detection. Singapore Management University Institutional Knowledge (InK) (Singapore Management University), 12173–12182. doi:10.1109/CVPR42600.2020.01219

Papamakarios, G., Nalisnick, E. T., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, *22*(57), 1–64. doi:10.48550/arXiv.1912.02762

Ran, Q., Liu, Z., Sun, X., Sun, X., Zhang, B., Guo, Q., & Wang, J. (2021). Anomaly detection for hyperspectral images based on improved low-rank and sparse representation and joint gaussian mixture distribution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 6339–6352. doi:10.1109/JSTARS.2021.3087588

Ristea, N.-C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., & Shah, M. (2022). Self-Supervised predictive convolutional attentive block for anomaly detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13576–13586. doi:10.1109/CVPR52688.2022.01321

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., & Gehler, P. V. (2022). Towards total recall in industrial anomaly detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14318–14328. doi:10.1109/CVPR52688.2022.01392

Rudolph, M., Wandt, B., & Rosenhahn, B. (2021). Same same but differNet: Semi-Supervised defect detection with normalizing flows. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 19007–11916. doi:10.1109/WACV48630.2021.00195

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., & Rabiee, H. R. (2021). Multiresolution Knowledge Distillation for Anomaly Detection. *Wiardi Beckman Foundation (Wiardi Beckman Foundation)*, 14902–14912. 10.1109/CVPR46437.2021.01466

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *Lecture Notes in Computer Science*, *10265*, 146–157. doi:10.1007/978-3-319-59050-9_12

Schwartz, E., Arbelle, A., Karlinsky, L., Harary, S., Scheidegger, F., Doveh, S., & Giryes, R. (2024). MAEDAY: MAE for few- and zero-shot AnomalY-Detection. *Computer Vision and Image Understanding*, *103958*, 103958. Advance online publication. doi:10.1016/j.cviu.2024.103958

Sheynin, S., Benaim, S., & Wolf, L. (2021). A Hierarchical Transformation-Discriminating Generative Model for Few Shot Anomaly Detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8495–8504. doi:10.1109/ICCV48922.2021.00838

Sun, Z., Wang, J., & Li, Y. (2023). RAMFAE: A novel unsupervised visual anomaly detection method based on autoencoder. *International Journal of Machine Learning and Cybernetics*, *15*(2), 355–369. doi:10.1007/s13042-023-01913-7

Wang, Y., Yuan, X., Lin, Y., Gu, J., & Zhang, M. (2023). A semi-supervised multi-scale deep adversarial model for fan anomaly detection. *IEEE Transactions on Consumer Electronics*, ●●●, 1–1. doi:10.1109/TCE.2023.3267077

Xie, G., Wang, J., Liu, J., Lyu, J., Liu, Y., Wang, C., Zheng, F., & Jin, Y. (2024). IM-IAD: Industrial Image Anomaly Detection Benchmark in Manufacturing. *IEEE Transactions on Cybernetics*, *54*(5), 1–14. doi:10.1109/TCYB.2024.3357213 PMID:38381632

Xu, Z., He, D., Vijayakumar, P., Gupta, B. B., & Shen, J. (2023). Certificateless public auditing scheme with data privacy and dynamics in group user model of cloud-assisted medical WSNs. *IEEE Journal of Biomedical and Health Informatics*, *27*(5), 2334–2344. doi:10.1109/JBHI.2021.3128775 PMID:34788225

Yi, J., & Yoon, S. (2021). Patch SVDD: Patch-Level SVDD for anomaly detection and segmentation. *Lecture Notes in Computer Science*, *12627*, 375–390. doi:10.1007/978-3-030-69544-6_23

Zavrtanik, V., Kristan, M., & Skočaj, D. (2021). Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8330-8339). doi:10.1109/ICCV48922.2021.00822

Zemmouchi-Ghomari, L. (2021). How industry 4.0 can benefit from semantic web technologies and artefacts. *International Journal of Software Science and Computational Intelligence*, *13*(4), 64–74. doi:10.4018/IJSSCI.2021100105

Zhou, Z., Li, Y., Li, J., Yu, K., Kou, G., Wang, M., & Gupta, B. B. (2023). GAN-Siamese network for cross-domain vehicle re-identification in intelligent transport systems. *IEEE Transactions on Network Science and Engineering*, *10*(5), 2779–2790. doi:10.1109/TNSE.2022.3199919

Zou, Y., Jeong, J., Pemula, L., Zhang, D., & Dabeer, O. (2022). SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. *Lecture Notes in Computer Science*, *13690*, 392–408. doi:10.1007/978-3-031-20056-4_23

Lei Zhang is currently a graduate student at Hubei University of Automotive Technology, specializing in the fields of Pattern Recognition and Artificial Intelligence. Her research is primarily focused on anomaly detection in small samples, exploring effective ways to identify and manage abnormal patterns within data, especially under conditions of limited sample sizes.

Lyu Chengzhi currently serves as a lecturer at Hubei University of Automotive Technology. He earned his Doctorate in Engineering from South China University of Technology in 2021. His research primarily focuses on the innovative intersection of Artificial Intelligence (AI) and image pattern recognition, contributing to the field with seven published academic and conference papers on the subject.

Ziheng Chen is a graduate student focusing on the fields of Pattern Recognition and Artificial Intelligence. His research is centered on anomaly detection based on teacher models, exploring how to identify anomalies in data with higher accuracy and efficiency by emulating the decision-making process of teacher models.

Shaokang Li is currently pursuing his graduate studies, with a focus on Pattern Recognition and Artificial Intelligence. His research primarily concentrates on monocular vision depth estimation, dedicated to exploring accurate depth information estimation of scenes from single images.

Bin Xia is a graduate student, with a focus on Pattern Recognition and Artificial Intelligence. His research is dedicated to 3D reconstruction, exploring the reconstruction of three-dimensional models from two-dimensional images through computer vision techniques.