Texture-AD: An Anomaly Detection Dataset and Benchmark for Real Algorithm Development

Tianwu Lei[1*], Bohan Wang[1*], Silin Chen[1], Shurong Cao[1], Ningmu Zou[1, 2†]

[1]School of Integrated Circuits, Nanjing University, Suzhou, China
[2]Interdisciplinary Research Center for Future Intelligent Chips (Chip-X), Nanjing University, Suzhou, China
tianwulei@smail.nju.edu.cn, bohanwang@smail.nju.edu.cn, silin.chen@smail.nju.edu.cn, shurongcao@smail.nju.edu.cn,
nzou@nju.edu.cn

Abstract

Anomaly detection is a crucial process in industrial manufacturing and has made significant advancements recently. However, there is a large variance between the data used in the development and the data collected by the production environment. Therefore, we present the Texture-AD benchmark based on representative texture-based anomaly detection to evaluate the effectiveness of unsupervised anomaly detection algorithms in real-world applications. This dataset includes images of 15 different cloth, 14 semiconductor wafers and 10 metal plates acquired under different optical schemes. In addition, it includes more than 10 different types of defects produced during real manufacturing processes, such as scratches, wrinkles, color variations and point defects, which are often more difficult to detect than existing datasets. All anomalous areas are provided with pixel-level annotations to facilitate comprehensive evaluation using anomaly detection models. Specifically, to adapt to diverse products in automated pipelines, we present a new evaluation method and results of baseline algorithms. The experimental results show that Texture-AD is a difficult challenge for state-of-the-art algorithms. To our knowledge, Texture-AD is the first dataset to be devoted to evaluating industrial defect detection algorithms in the real world. The dataset is available at https://XXX.

Introduction

Industrial inspection algorithms are typically developed and tested using collected data before deployment, for use in automated quality control equipment on production lines. In recent years, a variety of detection methods have developed for detecting an anomalous image region in image data through contemporary machine learning approaches. These methodologies have demonstrated promising results on established datasets. Present evaluation strategies typically entail integrating flawless production data of a single object category during the training stage and evaluating performance using data containing anomalies.

The acquisition of flawless production data has become more accessible when contrasted with defective data. However, a production line is often required to deal with various specifications of similar products, such as gray cloth, red cloth, mesh cloth, different types of wafers as well as black brushed metal plates, gold frosted metal plates, etc. While these different specifications share certain common features, they also present significant differences. Additionally, minor fluctuations in external conditions, such as lighting environment and camera settings, result in a data distribution after deployment that is unlikely to align with the data collected during the training phase. This situation places increased requirements on the robustness of the algorithms.

Humans have the natural ability to visually discern the similarities and differences in images and to detect defects and irregularities within them. Currently, there are many commonly used datasets for anomaly detection, which vary greatly in the scenes and scale they contain. For example, datasets related to cloth texture(Ninja 2024; Silvestre-Blanes et al. 2019) generally have a good amount of data, but they differ significantly from actual production scenarios. In addition, as chips become an increasingly important field of research worldwide, wafer defect detection has become an essential part of the process. Therefore, the demand for wafer defect detection datasets(Wu, Jang, and Chen 2015) in industrial inspection is also growing, yet there are very few open-source wafer defect detection datasets available. Moreover, there are more datasets related to metal defects in industrial production(Bao et al. 2021; Song, Song, and Yan 2020; Zhao et al. 2022; Niu et al. 2021; Feng, wen Gao, and Luo 2021; Zhang et al. 2021), but they generally include material types and apply to a more limited range of scenarios. There are also datasets related to crack defects(Guo et al. 2020; Xu et al. 2019), such as cracks in bridge surfaces and concrete floors.

So far, modern machine learning systems have encountered considerable challenges in addressing related issues, mainly because the existing datasets are not particularly well-suited to real-world scenarios. Currently, the evaluation of anomaly detection algorithms often relies on datasets such as MVTec(Bergmann et al. 2019), where the features of flawless and defective items show a high degree of consistency, leading to higher performance metrics than actual deployment. Therefore, this paper proposes the Texture-AD dataset(Texture-ad 2024), which clearly demonstrates the differences between Texture-AD and the MVTec dataset in Table 1. As shown in Figure 1, the training data provided by the MVTec dataset and the test data belong to completely
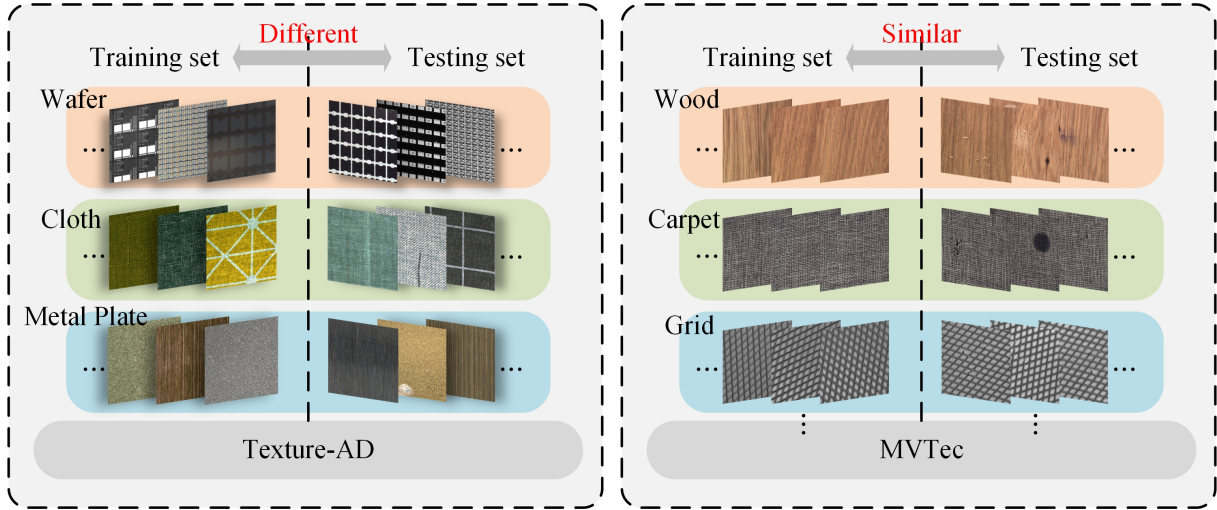

Figure 1: Difference between existing evaluation methods and actual situation

Table 1: Evaluation Protocol Difference Between Texture-AD and MVTec

Category	Train		Test	
	Images	Category Labels	Images	Category Labels
MVTec	O	O	O	O
Ours	O	O	O	X

the same product, making it impossible to correctly evaluate the algorithms under development. Therefore, in Texture-AD, we provide a variety of specifications of three products as the training set, and at the same time, provide the same type of products with different specifications from the train set as the test set, which can evaluate the performance of the algorithm based on the consideration of algorithm robustness and generalization ability. The training set of this dataset includes 15 subclasses of cloth images, 14 subclasses of wafer images and 10 subclasses of metal plate images. All cloth images come from the same type of cloth, wafer images come from 14 different subclasses of wafers and metal plate images come from metal plates with 5 different colors of brushed and matte surfaces, photographed under similar lighting conditions. The test set includes defective cloth images, wafer images and metal plate images photographed from the actual production process, which show slight differences in camera settings, lighting conditions and the design of cloth, wafers and metal plates compared to the training set.

The contributions of our paper can be summarized into three main aspects:

- We present a novel and comprehensive dataset for unsupervised anomaly detection in industrial quality inspection. It simulates real-world industrial inspection scenarios and it has a sufficient number of data samples and data scale, including 43120 high-resolution images collected in various optical environments from 39 different subclasses under three major categories, which contain a variety of different types of defects.

- We conduct a comprehensive evaluation of current state-of-the-art methods for unsupervised anomaly detection, assessing their segmentation and classification performance on the anomalous images during development process.

- We provide a well-designed evaluation protocol to compare the performance of unsupervised anomaly detection algorithms in actual development environments.

Related Work

Computer vision equipment for detecting surface defects has largely replaced manual inspections across industries like 3C electronics, automotive, machinery, semiconductors, chemicals and so on. Traditional methods use standard image processing and classifiers with handcrafted features, while effective imaging schemes ensure clear defect visibility under uniform lighting. Recently, deep learning has become prevalent for defect detection.

DAGM2007 dataset(Ninja 2024) is artificially generated but resembles real-world problems. Six categories referred to as the development dataset, should be used for algorithm development. The remaining four categories (referred to as the competition dataset) can be used to evaluate performance. AITEX dataset(Silvestre-Blanes et al. 2019) is an image dataset focused on the textile industry, designed to support research and application of machine learning and computer vision technology in the field of textile quality inspection. However, the aforementioned two datasets have issues with unclear defect labeling and a rather singular background type and defect type, which cannot fully simulate

the complex detection scenarios in actual industrial environments.

The WM-811K dataset(Wu, Jang, and Chen 2015) is a dataset specifically for semiconductor wafer map defect type identification, with images in the dataset mainly coming from actual production environments of wafer maps, obtained through electrical testing, and used to describe the state of wafer defects. However, the WM-811K represents without texture details and pattern information.

A dataset(Bao et al. 2021) collected six typical surface defects of hot-rolled steel strips. This surface defect dataset faces two major challenges: large differences in appearance among defects within the same category, and similarities between defects of different categories, with defect images affected by lighting and material changes. The NEU-surface-defect-database(Song, Song, and Yan 2020) has six typical surface defects of hot-rolled steel strips, namely rolling scale, patches, cracks, pitted surfaces, inclusions and scratches. The improved X-SDD dataset(Zhao et al. 2022) includes: seven typical types of hot-rolled steel strip defect images, due to the imbalance of sample quantity in X-SDD, it provides conditions for researchers to solve the problem of sample imbalance. The SD-saliency-900 dataset(Niu et al. 2021) includes three types of steel strip surface defects (inclusions, patches and scratches), including steel surface defect detection images and corresponding pixel-level binary masks. RSDDS-113 dataset(Feng, wen Gao, and Luo 2021), with samples taken from the actual industrial production line of a section steel factory, collects 20 track sections with defect information. Each pair of images in this dataset consists of a left camera image and the corresponding depth image; the dataset has a high degree of annotation credibility, but the amount of data samples is fewer. The Rail-5k dataset(Zhang et al. 2021) is used for the task of steel rail surface defect detection. The dataset can be used for two settings, the first is a supervised setting trained with marked images, the fine-grained nature of defect categories and long-tail distribution makes it difficult for visual algorithms to solve. The second is a semi-supervised learning setting promoted by unmarked images, including possible image damage and domain shift with marked images. The dataset can support both supervised and semi-supervised learning settings. In actual production, there may be unknown types of defects, making it difficult for the aforementioned traditional datasets based on known defect patterns to cope. In addition, it is difficult to obtain a large number of defect samples in the aforementioned datasets, leading to the problem of small sample sizes when training deep learning models.

The Concrete Crack Images for Classification dataset(Guo et al. 2020) is created specifically for the task of concrete crack classification. This dataset typically contains tens of thousands of images of concrete surfaces, showing cracks of different types and severities. The Crack-Detection dataset(Xu et al. 2019) is designed specifically for crack detection tasks, containing images for training and evaluating crack identification algorithms. These images usually come from various material surfaces, especially concrete and other construction engineering materials, because cracks in these materials may lead to structural

problems. The images in the aforementioned datasets have issues with varying quality, including resolution, lighting conditions, angles and background complexity, which may affect the performance of crack detection algorithms in the deployment process.

MVTec(Bergmann et al. 2019) contains images of anomalous samples with various defects, manually generated. This is a popular dataset for unsupervised anomaly detection that simulates real-world industrial inspection scenarios. The dataset provides the possibility of evaluating unsupervised anomaly detection methods for various textures and object classes with different types of anomalies. Since it provides pixel-level precise ground truth labels for the abnormal areas in the images, it is possible to evaluate anomaly detection methods for image-level classification and pixel-level segmentation.

In industrial settings, the prevalence of normal samples over defective ones creates a dataset imbalance, affecting model training and generalization. Acquiring a significant number of defective samples is costly and time-consuming, especially for rare defects. Current datasets may not cover all defect types, limiting the model's ability to identify unusual defects. The complexity of industrial products' appearance and potential labeling inconsistencies add to the challenge of defect detection. Moreover, the need for real-time responses in industry is often not met by existing datasets, leading to models that may not perform well in new environments.

Dataset

The anomaly detection dataset we propose includes 15 subclasses of cloth, covering a variety of colors, materials and texture defects, 14 different subclasses of wafers and 10 subclasses of metal plates, including 5 colors each with brushed and matte finishes, totaling 10 subclasses of textures. The defects in our dataset are imperfections that occur in actual production environments, making it extremely valuable for the study of industrial quality inspection algorithms. Cloth defects include pencil marks, cuts, marker stains, water stains, black and white dots, threads, inconsistent sewing distances and color differences caused by dyeing. Wafer and metal plate defects include scratches, stains and inherent manufacturing defects, all of which naturally occur in the production process. As shown in Figure 2, Our dataset contains a total of 43120 images, with 28973 images used for training and validation, and 14147 images for testing. The training set includes only defect-free images. The test set contains two types of images: images with various types of defects and defect-free images. Figure 3 shows the percentage of the image area occupied by the anomalous regions.

Specific to the division of the dataset, we provide good production images from multiple subclasses for each category as the training set, allowing the model to learn the characteristics and differences of each subclass. At the same time, we also provide defect images and good production images from the same category for the test set to evaluate the model's recognition ability when facing actual defects. The number of samples for each category and the specific allocation of subclasses are detailed in the appendix for reference.

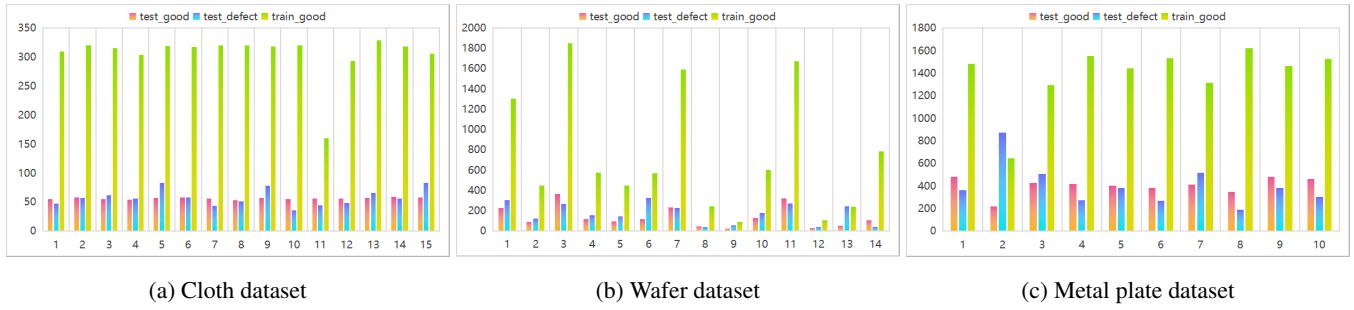(a) Cloth dataset   (b) Wafer dataset   (c) Metal plate dataset

Figure 2: Data Statistics (a)The cloth dataset consists of a total of 6283 images, with 4569 images in the training set and 1714 images in the test set. (b)The wafer dataset consists of a total of 14861 images, with 10525 images in the training set and 4336 images in the test set. (c)The metal plate dataset consists of a total of 21976 images, with 13879 images in the training set and 8097 images in the test set.
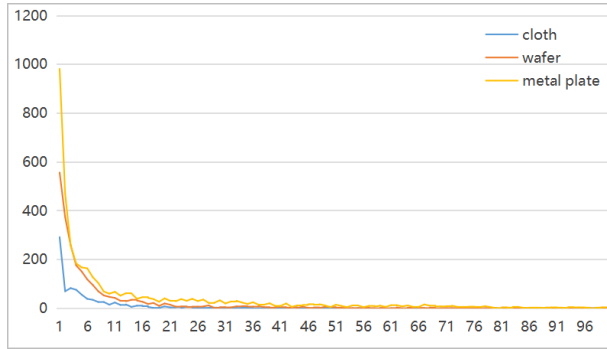

Figure 3: Statistics of the percentage of the image area occupied by the anomaly region

Data Generation

All images were captured using a high-resolution industrial camera (MV-CS200-10 GC) at a resolution of 5472×3648 pixels, in conjunction with two light sources. The optical scheme was altered by adjusting the position and brightness of the light sources. Our image acquisition and defect annotation process is depicted in Figure 4. The defects in our dataset were manually annotated using the Labelme annotation tool. To better align with the defects produced in the industrial manufacturing process, we created some artificial defects on the cloth, while the wafers and metal plates exhibited naturally occurring defects. Subsequently, these images were cropped to the appropriate output size. All images have a resolution of 1024×1024 pixels. The training set images were obtained under relatively stable lighting conditions. However, for the test set, we intentionally varied the optical scheme to simulate the imaging discrepancies between the algorithm training phase and actual deployment. We provided pixel-level ground truth annotations for each defective image area. The specific quantities for each category are listed in Figure 4.

Anomaly Detection Methods

The current research trend in anomaly detection is primarily focused on unsupervised anomaly detection. This trend has emerged due to the fact that obtaining anomalous samples requires a significant investment of human and financial resources. In this research context, training data contains only normal samples, while test data includes both normal and anomalous samples. Industrial image anomaly detection is a specific branch within the field of anomaly detection, and we mainly evaluate and compare it using the following three research directions.

Synthesis-based Anomaly Detection

Some supervised learning methods use a limited number of anomaly samples to synthesize more anomaly samples to enhance training effectiveness. For example, A basic architecture that integrates CycleGAN(Chu, Zhmoginov, and Sandler 2017) with ResNet/U-Net as the generator is used to transfer defects from one image to another(Rippel, Müller, and Merhof 2020). SDGAN(Liu, Wu, and Lv 2023) achieved better results than CycleGAN by improving the style transfer network. DRAEM(Zavrtanik, Kristan, and Skocaj 2021) first restores the normal image with pseudo-anomaly interference to obtain feature representation and then uses a discriminator network to distinguish anomalies, demonstrating excellent performance. Although this field has made certain research progress, it still has a huge development space compared to other fields with clear research directions.

Reconstruction-based Anomaly Detection

These methods are based on the assumption that a reconstruction model trained only on normal samples can successfully reconstruct images in normal areas(Bergmann et al. 2018; Chen et al. 2022; Zhang, Wang, and Kuo 2021; Sabokrou et al. 2018; You et al. 2022) but fail in abnormal areas. Early attempts included autoencoders(AE)(Bergmann et al. 2018; Collin and De Vleeschouwer 2021), variational autoencoders(VAE)(Zhang, Wang, and Kuo 2021; Kingma and Welling 2022) and generative adversarial networks(GAN)(Sabokrou et al. 2018; Akcay, Atapour-
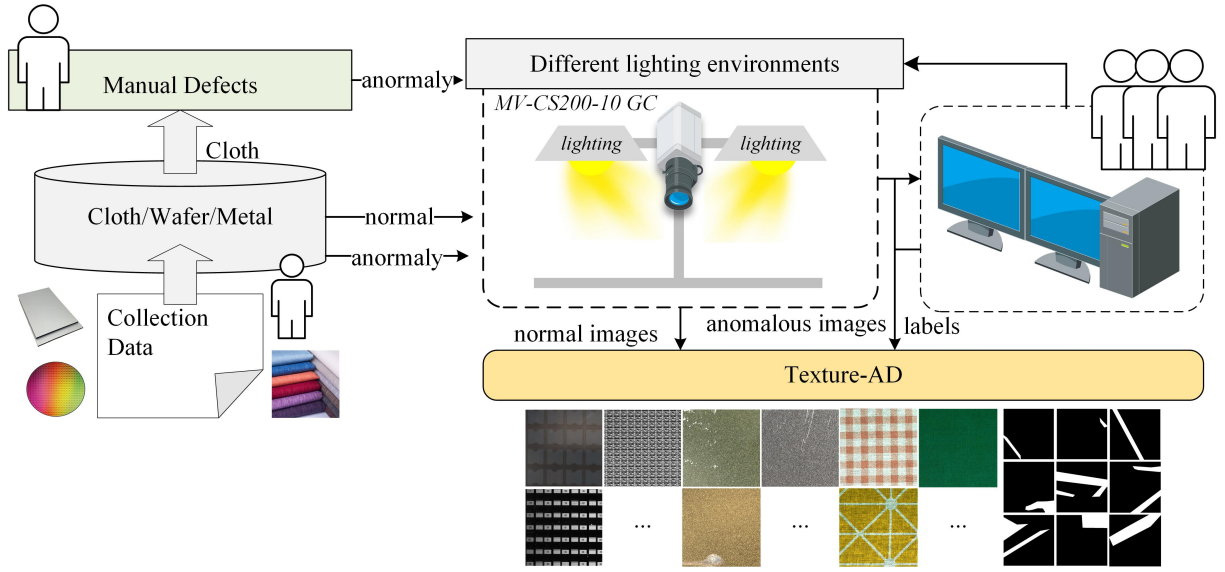
Figure 4: Image acquisition and defect annotation processes. The Texture-AD images were captured using a high-resolution industrial camera (MV-CS200-10 GC). The optical scheme was altered by adjusting the position of the light source and the brightness of two light sources. The cloth images include both artificial and natural defects, while the wafer and metal plate images consist solely of natural defects. The defect annotation work for the images was performed using Labelme.

Abarghouei, and Breckon 2018; Perera, Nallapati, and Xiang 2019; Zaheer et al. 2020). However, these methods may cause the model to learn certain tricks, leading to the effective recovery of anomalies as well. To address this issue, researchers have adopted various strategies, such as introducing guidance information (structure(Zhou et al. 2020) or semantics(Shi, Yang, and Qi 2021; Xia et al. 2020)), memory mechanisms(Gong et al. 2019; Hou et al. 2021; Park, Noh, and Ham 2020), iterative mechanisms(Dehaene et al. 2020), image masking strategies(Yan et al. 2021) and pseudo-anomaly(Collin and De Vleeschouwer 2021; Pourreza et al. 2020).PyramidFlow(Lei et al. 2023) based on the transformer and further design set a new record on MVTec.

Feature-Embedding Based Methods

Feature embedding methods are committed to distinguishing normal and abnormal samples at the feature representation level. Uniformed Students(Bergmann et al. 2020) pioneered the use of discriminative latent embeddings for anomaly detection. This model is simple and effective, significantly outperforming other benchmark methods. STPM(Wang et al. 2021) and MKD(Salehi et al. 2020) utilize multi-scale features on different network layers for feature distillation, although there are differences in their methods. In addition, SimpleNet(Liu et al. 2023) has achieved satisfactory results by introducing noise into the feature embedding to simulate negative samples.

Benchmark

Baseline Methods

SimpleNet SimpleNet(Liu et al. 2023) proposed a simple and easy-to-apply network for detecting and localizing

anomalies in images. We evaluated using the publicly available SimpleNet implementation on Pytorch. The backbone network used Wide Resnet50 as the backbone network, setting the feature dimension of the feature extractor to 1536 to accommodate 329×329 sized input images. The anomaly feature generator added isotropic Gaussian noise $N(0, \sigma^2)$, where σ defaults to 0.015. The subsequent discriminator includes a linear layer, batch normalization layer, leaky ReLU with a slope of 0.2 and a linear layer. The Adam optimizer was used, with learning rates of 0.0001 and 0.0002 set for the feature adapter and discriminator, respectively and a weight decay of 0.00001. Each dataset was trained for 160 epochs with a batch size of 8.

PyramidFlow PyramidFlow(Lei et al. 2023) proposed a new anomaly localization method, which is based on the defect contrastive localization paradigm using a pyramid of normalization flows for multi-scale fusion and volume normalization to achieve high-resolution defect localization. We used a fixed pyramid layer number $L = 8$, image resolution of 256×256 and channel number $C = 24$, and varied the stacked layer number D to explore the trends in memory usage and model parameterization. During training, sample mean normalization was used, and the running mean was updated with a momentum of 0.1. At test time, volume normalization was based on the running mean.

Mean-Shift Mean-Shift(Reiss and Hoshen 2022) introduced a novel self-supervised representation learning method to improve anomaly detection. It pointed out that traditional contrastive learning methods are not suitable for pre-trained features, hence they proposed the Mean-Shifted Contrastive Loss. In the experiments targeting ResNet152, we fine-tuned the last two blocks of a ResNet152 model

Table 2: Comparison of state-of-the-art works on the cloth of Texture-AD.Image-AUROC (top row) and Pixel-AUROC(bottom row) are displayed in each entry.

Category	subclass1	subclass2	subclass3	subclass4	subclass5	Average
SimpleNet	65.08	59.26	58.83	**70.40**	68.47	**64.41**
	58.30	51.52	**63.48**	**70.68**	54.47	59.69
PyramidFlow	57.88	63.18	60.74	59.39	49.72	58.18
	68.00	57.06	60.74	57.26	34.84	55.58
Mean-Shift	**66.22**	33.66	**66.21**	65.69	39.54	54.26
	-	-	-	-	-	-
DRAEM	57.58	50.21	55.44	58.01	55.95	55.44
	60.99	**65.36**	56.91	53.45	**77.03**	**62.75**
MSFlow	50.00	54.01	50.00	50.00	50.14	50.83
	56.11	63.14	51.66	47.44	42.23	52.12
EfficientAD	65.65	**76.98**	55.69	42.38	**72.20**	62.58
	62.76	58.92	47.08	38.75	61.77	53.86

Table 3: Comparison of state-of-the-art works on the wafer of Texture-AD.Image-AUROC (top row) and Pixel-AUROC(bottom row) are displayed in each entry.

Category	subclass1	subclass2	subclass3	subclass4	Average
SimpleNet	52.11	**59.66**	53.66	50.68	54.03
	57.18	**66.16**	**57.58**	**53.40**	**58.58**
PyramidFlow	55.54	43.35	52.76	46.36	49.50
	51.23	39.47	51.52	44.63	46.71
Mean-Shift	52.83	53.29	55.44	48.28	52.47
	-	-	-	-	-
DRAEM	**55.69**	57.09	**59.22**	**52.46**	**56.12**
	44.91	34.10	35.01	43.59	39.40
MSFlow	51.19	49.78	53.64	50.00	51.15
	44.91	34.10	35.01	43.59	39.40
EfficientAD	50.28	42.25	50.23	45.51	47.07
	55.76	33.98	51.53	40.02	45.32

Table 4: Comparison of state-of-the-art works on the metal plate of Texture-AD. Image-AUROC (top row) and Pixel-AUROC(bottom row) are displayed in each entry.

Category	subclass1	subclass2	subclass3	Average
SimpleNet	59.07	**59.87**	57.83	58.92
	62.27	**58.33**	58.97	59.86
PyramidFlow	52.87	48.74	58.92	53.51
	53.42	48.86	57.67	53.31
Mean-Shift	44.34	47.39	45.04	53.29
	-	-	-	-
DRAEM	52.07	56.32	51.48	45.59
	58.41	51.53	57.31	55.75
MSFlow	62.90	53.54	59.78	58.74
	65.37	57.34	**60.37**	**61.02**
EfficientAD	**65.27**	55.46	**68.73**	**63.30**
	59.69	51.04	54.91	55.21

pre-trained on the ImageNet dataset and added an ℓ_2 normalization layer, a process that lasted for 10 training epochs. For the experiments with ResNet18, we fine-tuned the entire backbone of a ResNet18 model pre-trained on ImageNet and similarly added an ℓ_2 normalization layer, a process that included 20 training epochs. In both cases, we minimized the Mean-Shifted Contrastive loss function with a temperature parameter τ set to 0.25. We used the Stochastic Gradient Descent (SGD) optimizer with a weight decay of 5×10^{-5}, and without momentum. We set the size of each mini-batch to 64.
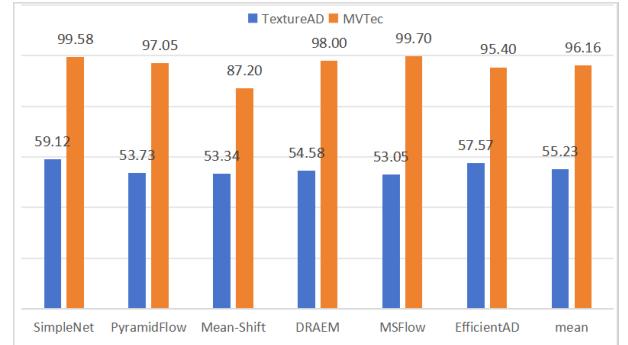

Figure 5: The comparison of the average Image-AUROC obtained by various algorithms on Texture-AD and MVTec

DRAEM In addition to reconstruction methods, DRAEM(Zavrtanik, Kristan, and Skocaj 2021) primarily regards surface anomaly detection as a discriminative problem and proposes a Discriminatively Trained Reconstruction Anomaly Embedding Model (DRAEM). This method learns the joint representation of anomalous images and their anomaly-free reconstructions while learning the decision boundary between normal and anomalous examples. The method can directly localize anomalies without the need for additional complex post-processing of the network output and can be trained using simple and universal anomaly simulation. In our experiments, the network was trained for 700 epochs. The learning rate was
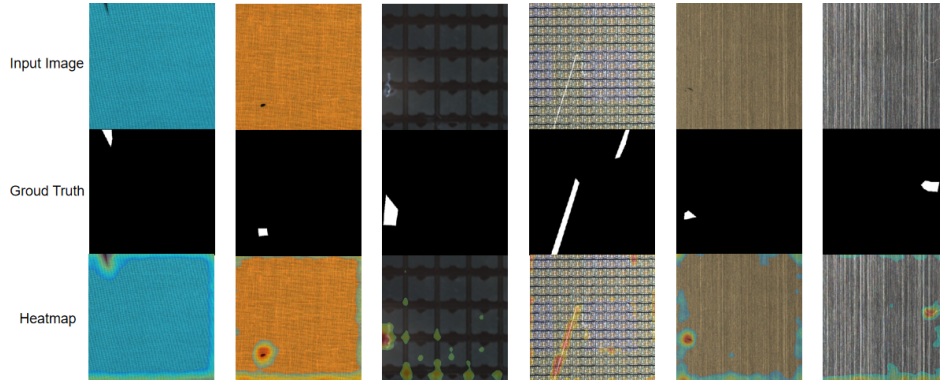
Figure 6: Visualization of SimpleNet results. It presents the anomaly segmentation results for three categories of materials in Texture-AD: cloth, wafer and metal plate. The top row demonstrates the origin image, the medium row shows pixel defect region annotation, and the bottom row is the heatmap of SimpleNet.

set to 10^{-4}, and it was multiplied by 0.1 after 400 and 600 epochs. Image rotation from -45 to 45 degrees was used as a data augmentation method.

MSFlow MSFlow(Zhou et al. 2023) proposed a multi-scale flow-based framework for unsupervised anomaly detection, which utilizes normalization flows to handle features at different scales to adapt to anomalies of various sizes. During the experimental process, we used Wide ResNet50 and ResNet18 as feature extractors. The training was conducted with a batch size of 16. The optimizer used was Adam with an initial learning rate of10^{-4}, and the learning rate was reduced at 70% and 90% of the training progress.

EfficientAD EfficientAD(Batzner, Heckler, and König 2024) proposed a lightweight feature extractor that processes images with millisecond-level latency on modern GPUs, using a student-teacher approach to detect anomalous features and effectively detect logical anomalies. In the experiments, we set the hard feature loss mining factor (phard) to 0.999, meaning that on average, 10% of the values in each dimension are used for backpropagation. The Adam optimizer was used with an initial learning rate of 10^{-4} and a weight decay of 10^{-5}. During training, if the number of iterations exceeded 66500, the learning rate was reduced to 10^{-5}.

Evaluation Method

Train and Test data As shown in Table 1, the information available during the training process is the same as for MVTec, but the sub-category labels cannot be used during the testing process.

Data Augmentation Since the evaluated methods based on deep learning are typically trained on large datasets, data augmentation is performed for these methods for both textures and objects. We resize the image to fit the shape of the model input. Additional mirroring is applied. We augment each category to create 10000 training images.

Evaluation Metric Following prior works(Bergmann et al. 2019; Zaheer et al. 2020; Bergmann et al. 2020), the Area Under the Receiver Operating Curve (AUROC)is used as the evaluation metric for anomaly detection. Image-level anomaly detection performance is measured via the standard Area Under the Receiver Operator Curve, which we denote as I-AUROC. For anomaly localization, we use an evaluation of pixel-wise AUROC (denoted as P-AUROC).

Result

As shown in Table 2, Table 3 and Table 4, we present the evaluation results of anomaly image classification and anomaly region segmentation for all methods and dataset categories, respectively. No method performs consistently well across all texture categories. In the cloth category, SimpleNet outperforms the other methods. But in the wafer category, DRAEM performs better than SimpleNet. In the metal plate category, EfficientAD leads the second place by 4.38% in I-AUROC. As shown in Figure 5, when applying our dataset Texture-AD for evaluation alongside the MVTec dataset, it was found that the evaluation results of our dataset are generally lower, which can expose the problem domains where the algorithm fails, facilitating targeted optimization of the algorithm's weak points in subsequent improvements. Here are the evaluation results of each method. Some examples of performance were provided.(Figure 6).All experimental results are the mean of 3 replicates.

Conclusion

We introduce the Texture-AD Anomaly Detection Benchmark, a novel dataset for unsupervised anomaly detection that mimics real-world industrial detection scenarios. The dataset provides a way to evaluate unsupervised anomaly detection methods in realistic algorithm development scenarios. Since pixel-accurate ground truth labels of anomaly regions in images are provided, both image-level classification and pixel-level segmentation anomaly detection methods can be evaluated. Several state-of-the-art methods are evaluated on this dataset. The evaluation provided a benchmark for showing how different algorithms perform in real-

world application scenarios and indicating that there is still much room for improvement. We hope that the proposed dataset will stimulate the development of new unsupervised anomaly detection methods.

References

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2018. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. arXiv:1805.06725.

Bao, Y.; Song, K.; Liu, J.; Wang, Y.; Yan, Y.; Yu, H.; and Li, X. 2021. Triplet-Graph Reasoning Network for Few-Shot Metal Generic Surface Defect Segmentation. *IEEE Transactions on Instrumentation and Measurement*, 70: 1–11.

Batzner, K.; Heckler, L.; and König, R. 2024. EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. arXiv:2303.14535.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9584–9592.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Bergmann, P.; Löwe, S.; Fauser, M.; Sattlegger, D.; and Steger, C. 2018. Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders. *CoRR*, abs/1807.02011.

Chen, L.; You, Z.; Zhang, N.; Xi, J.; and Le, X. 2022. UTRAD: Anomaly detection and localization with U-Transformer. *Neural Networks*, 147: 53–62.

Chu, C.; Zhmoginov, A.; and Sandler, M. 2017. CycleGAN, a Master of Steganography. *CoRR*, abs/1712.02950.

Collin, A.-S.; and De Vleeschouwer, C. 2021. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with Stain-shaped noise. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 7915–7922.

Dehaene, D.; Frigo, O.; Combrexelle, S.; and Eline, P. 2020. Iterative energy-based projection on a normal data manifold for anomaly localization. arXiv:2002.03734.

Feng, X.; wen Gao, X.; and Luo, L. 2021. X-SDD: A New Benchmark for Hot Rolled Steel Strip Surface Defects Detection. *Symmetry*, 13: 706.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and van den Hengel, A. 2019. Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection. arXiv:1904.02639.

Guo, L.; Li, R.; Jiang, B.; and Shen, X. 2020. Automatic crack distress classification from concrete surface images using a novel deep-width network architecture. *Neurocomputing*, 397: 383–392.

Hou, J.; Zhang, Y.; Zhong, Q.; Xie, D.; Pu, S.; and Zhou, H. 2021. Divide-and-Assemble: Learning Block-wise Memory for Unsupervised Anomaly Detection. arXiv:2107.13118.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Lei, J.; Hu, X.; Wang, Y.; and Liu, D. 2023. Pyramid-Flow: High-Resolution Defect Contrastive Localization using Pyramid Normalizing Flow. arXiv:2303.02595.

Liu, Y.; Wu, G.; and Lv, Z. 2023. SDGAN: A novel spatial deformable generative adversarial network for low-dose CT image reconstruction. *Displays*, 78: 102405.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023. SimpleNet: A Simple Network for Image Anomaly Detection and Localization. arXiv:2303.15140.

Ninja, D. 2024. Visualization Tools for Industrial Optical Inspection Dataset. https://datasetninja.com/industrial-optical-inspection. Visited on 2024-08-14.

Niu, M.; Song, K.; Huang, L.; Wang, Q.; Yan, Y.; and Meng, Q. 2021. Unsupervised Saliency Detection of Rail Surface Defects Using Stereoscopic Images. *IEEE Transactions on Industrial Informatics*, 17(3): 2271–2281.

Park, H.; Noh, J.; and Ham, B. 2020. Learning Memory-guided Normality for Anomaly Detection. arXiv:2003.13228.

Perera, P.; Nallapati, R.; and Xiang, B. 2019. OCGAN: One-Class Novelty Detection Using GANs With Constrained Latent Representations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2893–2901.

Pourreza, M.; Mohammadi, B.; Khaki, M.; Bouindour, S.; Snoussi, H.; and Sabokrou, M. 2020. G2D: Generate to Detect Anomaly. arXiv:2006.11629.

Reiss, T.; and Hoshen, Y. 2022. Mean-Shifted Contrastive Loss for Anomaly Detection. arXiv:2106.03844.

Rippel, O.; Müller, M.; and Merhof, D. 2020. GAN-based Defect Synthesis for Anomaly Detection in Fabrics. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, volume 1, 534–540.

Sabokrou, M.; Khalooei, M.; Fathy, M.; and Adeli, E. 2018. Adversarially Learned One-Class Classifier for Novelty Detection. *CoRR*, abs/1802.09088.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2020. Multiresolution Knowledge Distillation for Anomaly Detection. arXiv:2011.11108.

Shi, Y.; Yang, J.; and Qi, Z. 2021. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing*, 424: 9–22.

Silvestre-Blanes, J.; Albero-Albero, T.; Miralles, I.; Pérez-Llorens, R.; and Moreno, J. 2019. A Public Fabric Database for Defect Detection Methods and Results. *Autex Research Journal*, 19.

Song, G.; Song, K.; and Yan, Y. 2020. Saliency detection for strip steel surface defects using multiple constraints and improved texture features. *Optics and Lasers in Engineering*, 128: 106000.

Texture-ad. 2024. Texture-AD-Benchmark. https://huggingface.co/datasets/texture-ad/Texture-AD-Benchmark. Accessed: 2024-08-15.

Wang, G.; Han, S.; Ding, E.; and Huang, D. 2021. Student-Teacher Feature Pyramid Matching for Anomaly Detection. arXiv:2103.04257.

Wu, M.-J.; Jang, J.-S. R.; and Chen, J.-L. 2015. Wafer Map Failure Pattern Recognition and Similarity Ranking for Large-Scale Data Sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1): 1–12.

Xia, Y.; Zhang, Y.; Liu, F.; Shen, W.; and Yuille, A. 2020. Synthesize then Compare: Detecting Failures and Anomalies for Semantic Segmentation. arXiv:2003.08440.

Xu, H.; Su, X.; Wang, Y.; Cai, H.; Cui, K.; and Chen, X. 2019. Automatic Bridge Crack Detection Using a Convolutional Neural Network. *Applied Sciences*, 9: 2867.

Yan, X.; Zhang, H.; Xu, X.; Hu, X.; and Heng, P.-A. 2021. Learning Semantic Context from Normal Samples for Unsupervised Anomaly Detection. In *AAAI Conference on Artificial Intelligence*.

You, Z.; Yang, K.; Luo, W.; Cui, L.; Zheng, Y.; and Le, X. 2022. ADTR: Anomaly Detection Transformer with Feature Reconstruction. arXiv:2209.01816.

Zaheer, M. Z.; ha Lee, J.; Astrid, M.; and Lee, S.-I. 2020. Old is Gold: Redefining the Adversarially Learned One-Class Classifier Training Paradigm. arXiv:2004.07657.

Zavrtanik, V.; Kristan, M.; and Skocaj, D. 2021. DRÆM - A discriminatively trained reconstruction embedding for surface anomaly detection. *CoRR*, abs/2108.07610.

Zhang, K.; Wang, B.; and Kuo, C. J. 2021. PEDENet: Image Anomaly Localization via Patch Embedding and Density Estimation. *CoRR*, abs/2110.15525.

Zhang, Z.; Yu, S.; Yang, S.; Zhou, Y.; and Zhao, B. 2021. Rail-5k: a Real-World Dataset for Rail Surface Defects Detection. *CoRR*, abs/2106.14366.

Zhao, X.; Shi, J.; Yin, Q.; Dong, Z.; Zhang, Y.; Kang, L.; Yu, Q.; Chen, C.; Li, J.; Liu, X.; and Zhang, K. 2022. Controllable synthesis of high-quality two-dimensional tellurium by a facile chemical vapor transport strategy. *iScience*, 25(1): 103594.

Zhou, K.; Xiao, Y.; Yang, J.; Cheng, J.; Liu, W.; Luo, W.; Gu, Z.; Liu, J.; and Gao, S. 2020. Encoding Structure-Texture Relation with P-Net for Anomaly Detection in Retinal Images. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 360–377. Cham: Springer International Publishing. ISBN 978-3-030-58565-5.

Zhou, Y.; Xu, X.; Song, J.; Shen, F.; and Shen, H. T. 2023. MSFlow: Multi-Scale Flow-based Framework for Unsupervised Anomaly Detection. arXiv:2308.15300.