https://doi.org/10.37188/lam.2025.032

Article Open Access

MT-Former: Multi-Task Hybrid Transformer and Deep Support Vector Data Description to Detect Novel anomalies during Semiconductor Manufacturing

Hyunsu Jeong[1], Chiho Yoon[1], Hyunseok Lim[1], Jaesuk Chang[2], Sampa Misra[1] and Chulhong Kim[1,*]

Abstract

Defect inspection is critical in semiconductor manufacturing for product quality improvement at reduced production costs. A whole new manufacturing process is often associated with a new set of defects that can cause serious damage to the manufacturing system. Therefore, classifying existing defects and new defects provides crucial clues to fix the issue in the newly introduced manufacturing process. We present a multi-task hybrid transformer (MT-former) that distinguishes novel defects from the known defects in electron microscope images of semiconductors. MT-former consists of upstream and downstream training stages. In the upstream stage, an encoder of a hybrid transformer is trained by solving both classification and reconstruction tasks for the existing defects. In the downstream stage, the shared encoder is fine-tuned by simultaneously learning the classification as well as a deep support vector domain description (Deep-SVDD) to detect the new defects among the existing ones. With focal loss, we also design a hybrid-transformer using convolutional and an efficient self-attention module. Our model is evaluated on real-world data from SK Hynix and on publicly available data from magnetic tile defects and HAM10000. For SK Hynix data, MT-former achieved higher AUC as compared with a Deep-SVDD model, by 8.19% for anomaly detection and by 9.59% for classifying the existing classes. Furthermore, the best AUC (magnetic tile defect 67.9%, HAM10000 70.73%) on the public dataset achieved with the proposed model implies that MT-former would be a useful model for classifying the new types of defects from the existing ones.

Keywords: Semiconductor defect inspection, Deep-SVDD, Multi-task learning, Anomaly detection, Hybrid-transformer

Introduction

Increasing demand for high-speed devices accelerates new semiconductor nanofabrication technologies to embed more memory into small devices[1]. To ensure the quality of products from the nano-technology fabrications, using a scanning electron microscope (SEM) is required[2,3]. Defect inspection is critical in semiconductor wafer fabrication process for product quality improvement at reduced

Correspondence: Chulhong Kim (chulhong@postech.edu)
[1]Graduate School of Artificial Intelligence (GSAI), Department of Electrical Engineering, Convergence IT Engineering, Mechanical Engineering, Medical Science and Engineering, and Medical Device Innovation Center, Pohang University of Science and Technology (POSTECH), Pohang, South Korea
[2]Quality Intelligence System Team, SK Hynix, Icheon, 17336, South Korea
These authors contributed equally: Hyunsu Jeong, Chiho Yoon

production costs. In general, human inspection is time-consuming, labor-intensive and error-prone. Therefore, algorithmic methods for anomaly detection are increasingly essential[4–7].The rapid development of deep learning (DL) in computer vision[8–18] and the large amount of collected data in modern industry have made data-driven methods appealing to inspection engineers[19].

Deep autoencoder (DAE) is one of data-driven approaches to detect anomalies[20]. DAE creates the latent vectors from high-dimensional data, then reconstructs the data from the embedding vectors. When trained on only normal data, a DAE reconstructs normal cases well, meaning the reconstruction error $\|x-\hat{x}\|$ between the original image x and the reconstructed image $\hat{x}$ is small for normal cases. On the other hand, the reconstruction error of the model is higher for abnormal cases[21–27]. However, the DAE approach in semiconductor field has its limitation because of some defects that can be classified as normal owing to chip designing. Therefore, reconstruction loss function can be challenging for classifying semiconductor defects.

Generative adversarial networks (GANs)[28] are widely used for DL image analysis[29–32] and have also been adopted in the field of anomaly detection. AnoGAN[33] detects outliers by using a GAN that is trained on normal images. AnoGAN consists of two loss parts: residual loss and discrimination loss. In residual loss $L_R(z)$, a generator G usually creates normal images $G(z)$ from latent vector z, so an image error $\|\mathrm{x}-G(\mathrm{z})\|$ will be large if real data x is abnormal. The discrimination loss $L_D(z)$ involves training a discriminator f on normal images, ensuring that the latent vectors $f(x)$ of real normal images closely match the latent vectors $f(G(z))$ of generated images. This similarity is enforced using the loss function $\|f(x)-f(G(\mathrm{z}))\|$. Based on the two losses, the anomaly score, i.e. $(1-\lambda)\cdot L_R+\lambda\cdot L_D$, is utilized to identify anomaly cases[34]. GANomaly[35] makes AnoGAN's basic procedure more efficient by simultaneously generating images and encoding latent representation. GANomaly consists of a convolutional neural network E and a U-shaped generator G with an encoder G_E and a decoder G_D. The model uses the encoder loss to compute the anomaly score $|G_E(x)-E(G(x))\|$.

Deep Support Vector Domain Description (Deep-SVDD)[36] is a one-class classifier designed to train a model for anomaly detection-based objectives. Deep-SVDD is devised to find abnormal cases that are beyond the range of normality. As faulty anomaly data accumulates, supervised learning-based research has also been conducted to classify defect types[37,38]. Multiview data novelty detection using deep autoencoding support vector data descriptions (DMVSVDD) which jointly trains multiple DAEs to capture the correlated knowledge of multi-view data[39]. And multi-task learning strategy has also been used with Deep-SVDD, which simultaneously learns one-class classification objective in addition to using DAEs to learn reconstruction[40]. AdaDL-SVDD is trained on both normal and abnormal data to generate sparse representations with dictionary learning and adaptive boosting using some weak classifiers[41], while Deep-SVDD using VAEs improves anomaly detection accuracy by ensuring distinctly separate latent representations in DAE[42]. The Patch-SVDD that use patches has advantages because it performs anomaly detection on a per-patch basis. Since they use small patches for evaluation, they localize anomalous regions well[43,44].

Semiconductor-fabrication field is using DL models to analyze defects in wafers. To determine a specified cause of problems in the manufacturing process, many researchers try to distinguish wafer map patterns and defect types in SEM images[38,45–48]. Recently, SEM inspection needs anomaly-detection models that detect new defects that are different from the existing ones (Fig. 1). Many researchers developed anomaly models to distinguish defect types from normal cases[49–53], but it is difficult to

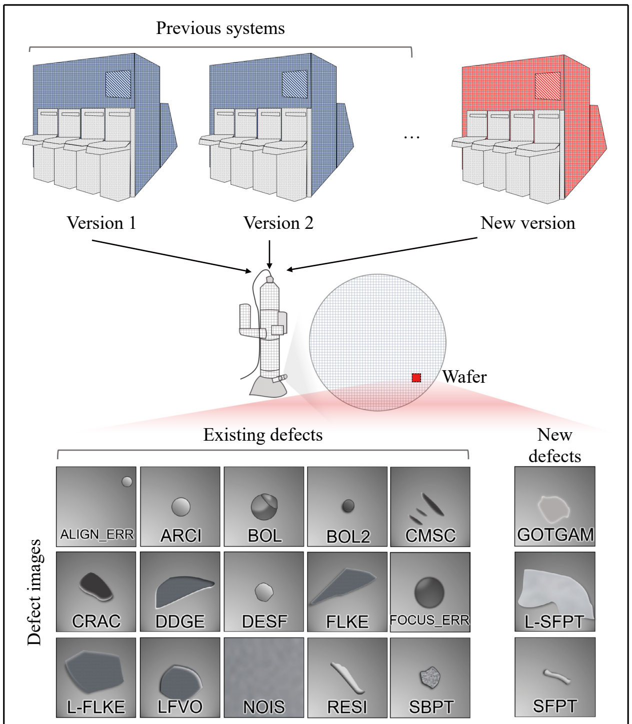

Fig. 1 Different defect types from various fabrication processes. New manufacturing technology can cause abnormal cases, which can be observed using scanning electron microscopy. In deep support vector domain description, existing defects are considered as normal cases and new types of defects as abnormal.

identify new defects due to numerous defect types. To enable analysis of the issue arising from the new processes, the new types of anomalies must be detected. In addition, an anomaly-detection model should consider irregular shapes and different sizes of defects from a complex fab process[54]. Therefore, a DL architecture is needed to extract appropriate features for this purpose.

Herein, we propose an MT-former framework that exploits multi-tasking learning (MTL) and efficient self-attention to distinguish new semiconductor defects from the existing defect types. The main contributions are as follows:

- Unlike conventional anomaly models that just distinguish defects from normal cases, the MT-former proposes to distinguish new defects from the existing defects.
- MTL stabilizes SVDD training process by mitigating the fluctuation of false positive rate and false negative rate. In addition, MTL achieves efficient clustering of each normal class, and thereby overcomes the limitations of the basic Deep-SVDD model, which indiscriminately clusters the existing defect types with various patterns into a single class.
- Hybrid transformer considering long-range dependencies improve results by identifying irregular patterns on a wide range of defects, and efficiently exploiting focal loss to successfully process imbalanced data (Fig. 1).
- In classification of defect types, a shared encoder from MT-former is also superior to a CNN model that simply learns defect classification from scratch.

Methods

Deep-SVDD

Deep-SVDD (Fig. 2a) is designed to map normal data into a hypersphere of minimum volume for one-class classification. A shared encoder learns transformation $\varphi(\cdot;W): X \rightarrow F$ for input data $X \in \mathbb{R}^{h \times w \times k}$ and an output feature $F \in \mathbb{R}^{1 \times 1 \times k'}$. The encoder includes $L \in \mathbb{N}$ layers that have weights $W = \{W^1, W^2, ..., W^L\}$. We basically adopt the upstream and downstream stage of the basic Deep-SVDD, but the classification task is simultaneously trained at the two stages.

During the upstream stage, DAE is used both to initialize weights of the shared encoder and to find a center point $c \in F$. DAE generates reconstruction images $\hat{X}$ and searches for optimal W^* to minimize a reconstruction error. Given training dataset $X = \{x_1, \ldots, x_N\}$ and $N \in \mathbb{N}$, the error is defined as:

$$L_{DAE} = \sum_{i=1}^{N} \|\hat{x}_i - x_i\|^2 + \frac{\lambda}{2} \sum_{l=1}^{2L} \|W^l\|^2 \quad (1)$$

The center point is calculated as an average of all

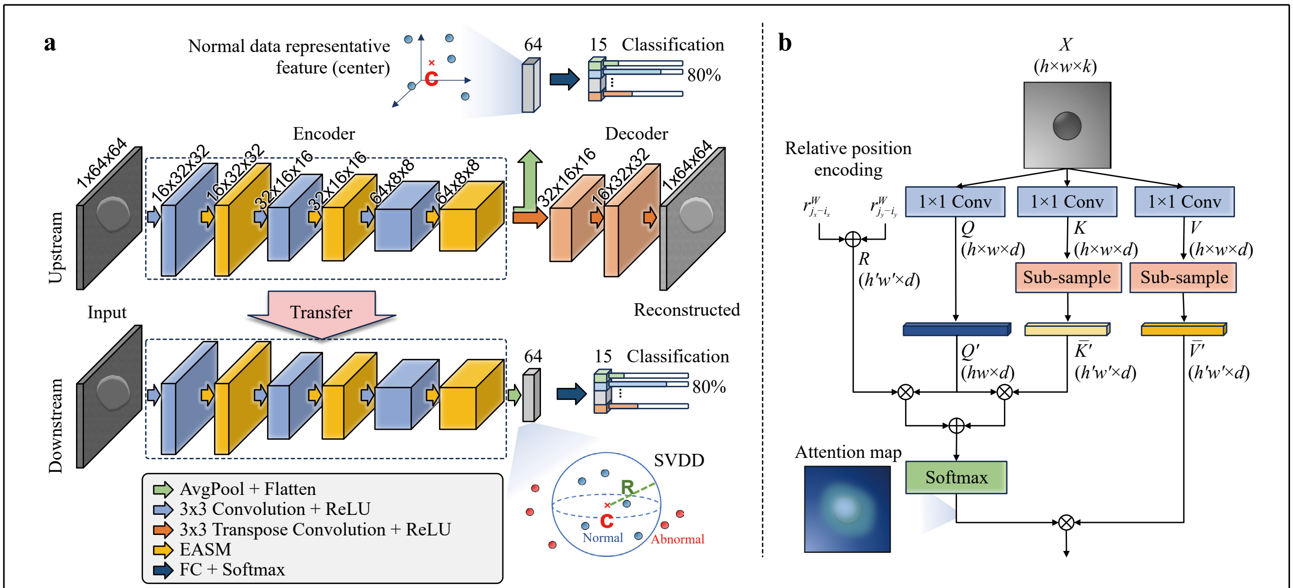

Fig. 2 Schematic overview of the MT-former framework. **a** The MT-former consists of an upstream and a downstream stage. At the upstream, U-shaped model with an encoder and a decoder is trained to reconstruct the existing defects. The encoder part simultaneously learns classification on the types of existing defect. At downstream, the encoder is fine-tuned by learning not only to categorize the existing defects cases, but also to cluster them. The encoder part is composed of the convolution and the efficient self-attention layer to extract both local and global features. **b** The efficient self-attention reduces the size of feature map on key and value.

features' coordinates as:

$$c = \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i; W^*) \tag{2}$$

Downstream is a second training phase, where the normal samples are clustered to the center point. The goal of the SVDD is to form the smallest hypersphere with radius R that encompasses normal samples from center point c.

$$L_{SVDD} = \sum_{i=1}^{N}\|\varphi(x_i; W) - c\|^2 + \frac{\lambda}{2}\sum_{l=1}^{L}\|W^l\|^2,\ for\ \forall_i \in [1, N] \tag{3}$$

$$s.t.\ \|\varphi(x_i; W) - c\|^2 \leq R^2,\ R > 0$$

Anomaly cases are identified if anomaly score $s(x)$ exceeds R, i.e. $s(x) > R^2$:

$$s(x) = \|\varphi(x; W) - c\|^2 \tag{4}$$

Hybrid Transformer Model

Models that apply transformers use multi-head self-attention (MHSA) modules to capture long-range dependency at different scales[55]. Such models require a huge training dataset, so a hybrid-transformer model, UT-Net, was introduced; it has an appropriate mix of convolutional layers and transformers[56]. We utilize the UT-Net's encoder which only applies convolution to the input image size. Unlike the encoder of UTNet, the self-attention module is applied on the input image size because our input image size is small.

UT-Net includes an efficient self-attention module (ESAM) to avoid inefficient and redundant computations. Given an input data $X \in \mathbb{R}^{h\times w\times k}$, 1×1 convolutions project it to the vectors that consist of query, key, value: $Q, K, V \in \mathbb{R}^{h\times w\times d}$. Most of the informative features in self-attention are contained in the largest singular values[57], K and V are downscaled to $\overline{K}, \overline{V} \in \mathbb{R}^{h'\times w'\times d}$ by sub-sample operation (Fig. 2b). $Q, \overline{K}, \overline{V}$ is sequentially flattened and transposed to $Q' \in \mathbb{R}^{o\times d}$ and $\overline{K'}, \overline{V'} \in \mathbb{R}^{u\times d}$ where $o = h\times w, u = h'\times w'$, and $u \ll o$, which is followed by a scaled dot-product defined as:

$$Attention\left(Q', \overline{K'}, \overline{V'}\right) = \underbrace{softmax\left(\frac{Q'\overline{K'}^T}{\sqrt{d}}\right)}_{o\times u}\underbrace{\overline{V'}}_{u\times d} \tag{5}$$

In addition, relative positional encodings use independent relative height and relative width information as self-attention augmentations, which prevent perturbation equivariance while allowing for translation equivariance[58,59]. The relative positional embeddings $r^W_{j_x-i_x}$ and $r^H_{j_y-i_y}$ between pixel $i = (i_x, i_y)$ and pixel $j = (j_x, j_y)$ are learned for relative width $j_x - i_x$ and height $j_y - i_y$. The relative attention logit for the strength of the relationship between pixel i and to pixel j is computed as:

$$l_{i,j} = \frac{q_i^T}{\sqrt{d}}\left(k_j + r^W_{j_x-i_x} + r^H_{j_y-i_y}\right) \tag{6}$$

where q_i is the i-th row of Q' and k_j is the j-th row of K'. The final self-attention formula is defined as:

$$Attention\left(Q', \overline{K'}, \overline{V'}\right) = \underbrace{softmax\left(\frac{Q'\overline{K'}^T + S^{rel}_W + S^{rel}_H}{\sqrt{d}}\right)}_{o\times u}\underbrace{\overline{V'}}_{u\times d} \tag{7}$$

where $S^{rel}_H[i, j] = q_i^T r^H_{j_y-i_y}$ and $S^{rel}_W[i, j] = q_i^T r^W_{j_x-i_x}$ are matrices of relative position logits, with $S^{rel}_H, S^{rel}_W \in \mathbb{R}^{hw\times h'w'}$ (Fig. 2).

Proposed MT-former

Our proposed method, MT-former (Fig. 2, Table 1), is composed of MTL and the hybrid transformer. MTL is a process of training a DL network on several related tasks at once with the intention that the shared knowledge learned from one task will increase accuracy on other tasks[60]. The goal of our task is to distinguish new defects from existing defects, which are considered to be 'normal' cases in this one-class classification study. However, we assumed that due to the variety of defect types in shape, normal clustering would not be successful if all of the existing defects were grouped into a single class. This is because considering normal classes with various patterns results in a broad data distribution and the range of clustering. The broad distribution makes it easier for the new defect data to be included in the distribution of the existing defects. Therefore, we suppose that MTL, by classifying the existing defect types and clustering them, enables the model to properly cluster the various-normal cases.

In this study, we take two MTL stages, one upstream and one downstream. During the upstream stage L_{Up}, reconstruction loss L_{DAE} and classification loss L_{Focal} are used. Focal loss L_{Focal} is applied to cover the existing defect classes imbalance problem[61]. p_t is the probability of classification by softmax when the model performs classification task on the existing defects. When focusing parameter γ increases, the learning weight of hard example with low probability also gets high. After initialization of the hybrid-transformed encoder, then during the downstream stage, the MT-former L_{Down} is proposed by simultaneously learning Deep-SVDD objective L_{SVDD} and the classification L_{Focal} on the hybrid-transformed encoder. The losses are defined as:

Table 1 Details of the architectures used in MT-former.

	Layer	Number of filters/heads	Filter size	Activation function
Attention encoder	Convolution (s = 2)	16	3*3	ReLU
	ESAM	3	–	–
	Convolution (s = 2)	32	3*3	ReLU
	ESAM	3	–	–
	Convolution (s = 2)	64	3*3	ReLU
	ESAM	3	–	–
Feature extraction	AdaptiveAvgPool2D	–	–	–
	Flatten	64	–	–
Classification	Fully Connected Layer	15	–	Softmax
Decoder	Conv2DTranspose (s = 2)	32	3*3	Relu
	Conv2DTranspose (s = 2)	16	3*3	Relu
	Conv2DTranspose (s = 2)	1	3*3	Tanh

$$L_{Focal} = -(1-p_t)^{\gamma}\log(p_t) \tag{8}$$

$$L_{Up} = L_{DAE} + L_{Focal} \tag{9}$$

$$L_{Down} = L_{SVDD} + L_{Focal} \tag{10}$$

In a nutshell, the proposed method is trained on only the existing defect-labeled dataset. Based on the output value, the model detects the new defect.

Experiments

Comparison models

We compared the accuracy of MT-former with six state-of-the-art deep anomaly detection models, i.e. DAE[20], GANomaly[35], AnoGAN[33], Patch-SVDD[43], Patchcore[44] and Deep-SVDD[36].

1. DAE[20] uses reconstruction error as a criterion for judging anomaly scores.
2. AnoGAN[33] is an early anomaly detection model based on GAN, which calculates anomaly scores by considering latent space in image space.
3. GANomaly[35] is a form in which an encoder is added to AnoGAN, and is more intuitive than AnoGAN to learn image and latent space at once.
4. Patch-SVDD[43] is an approach that can utilize local information by embedding in patch units. For comparison, Patch-SVDD's patch size is set to 32, which is half the input size.
5. Patchcore[44] extracts patch-wise features based on a model trained on ImageNet data to detect anomalies.
6. Deep-SVDD[36] obtains a hypersphere surrounding normal data, then uses it to identify abnormalities.

Evaluation metrics

To evaluate anomaly-detection accuracy, we defined abnormal cases as ‘positive’ and used four evaluation metrics: true positive rate (TPR, recall), false positive rate (FPR), signal-to-background ratio (S/B), and area under the receiver operating characteristic curve (AUC). Receiver operating characteristic (ROC) curves are used to visualize the tradeoff between TPR and an FPR at different thresholds, while AUC shows the overall detection accuracy as the area under the ROC curve. In classification experiments, we adopted weighted average AUC because from the viewpoints of quality inspection and costs, frequent defects are the most important. The metrics are defined as:

$$TPR = Recall = \frac{TP}{TP+FN} \tag{11}$$

$$FPR = \frac{FP}{FP+TN} \tag{12}$$

$$S/B = \frac{TP}{FP} \tag{13}$$

$$Weighted\ Average = \sum_{i=1}^{c} w_i \times s_i \tag{14}$$

where w_i is the number of data belonging to class i and s_i is the score of class i.

Data Acquisition

In this study, all experiments were conducted on 6078 datasets, including 24 new defect datasets from a domestic SK Hynix’s FAB process (SK-defect, SK Hynix, South Korea). SK data were collected in different settings and system environments. The defect images were provided in

64 to 80 pixel sizes, a size that allows engineers to visually identify defects and enables rapid analysis. The images typically have a field of view (FOV) ranging from 1 μm to 2 μm, which corresponds to approximately 15.6 nm to 31.3 nm per pixel for 64 × 64 pixel images. For each manufacturing process, defect types were defined considering image shape and process characteristics (Fig. 1). For instance, the ALIGN_ERR defect tends to exhibit alignment errors where the target is far from the center, the BOL defect often shows a round or circular shape, FLKE defect looks like flakes, which are similar to thin chip-like fragments[62], and the FOCUS_ERR is characterized by blurry or out-of-focus patterns in the imaging. The collected data were divided into 2951 training and 3127 testing (Table 2a). The 18 defect classes include three new types of defects and 15 existing types of defects. This study considers the existing defects as 'normal' cases for the Deep-SVDD task. The data is not publicly available. However, the authors will make the data available upon reasonable request and with the permission of SK Hynix.

For robustly the capability of our method, we evaluated our models on two public dataset. The magnetic tile defect dataset was previously utilized for another validation[63]. This dataset contains one non-defect case and five defect cases: blowhole, crack, fray, break, and uneven. To evaluate the ability to distinguish new defects, we performed validation by excluding the non-defect case and using only five defect classes, and set the 'fray' class that had the fewest instances as the 'new' defect. For normal defect data, we split the data 8:2 for training and testing, and all new defects were used for testing. In the final dataset, the number of training data was 287 and the number of testing data was 105, including 32 abnormal cases (Table 2b). The well-known HAM10000[64] consists of 6 subclasses, with an imbalanced dataset. The HAM10000 dataset is divided into 7007 training, 1003 validation, and 2005 testing (Table 2c). In our experiments with this dataset, we treated each subclass as a 'new' defect in turn, while considering the remaining subclasses as normal defects, which allowed us to observe the performance differences across classes.

Table 2 Class ditribution of train and test in defect image dataset.

(a) SK-defect

Data subclass	Train	Test		Total
		Normal	Abnormal	
ALIGN_ERR	178	90	–	268
ARCI	76	3	–	79
BOL	347	56	–	403
BOL2	278	1522	–	1800
CMSC	199	27	–	226
CRAC	257	75	–	332
DDGG	451	743	–	1194
DESF	222	296	–	518
FLKE	153	8	–	161
FOCUS_ERR	147	26	–	173
L_FLKE	26	2	–	28
LFVO	106	9	–	115
NOIS	150	60	–	210
RESI	119	12	–	131
SBPT	242	174	–	416
GOTGAM	–	–	9	9
L_SFPT	–	–	2	2
SFPT	–	–	13	13
Total	2951	3103	24	

(b) Magnetic tile defect

Data subclass	Train	Test		Total
		Normal	Abnormal	
Blowhole	92	23	–	115
Break	68	17	–	85
Crack	45	12	–	57
Uneven	82	21	–	103
Fray	–	–	32	32
Total	287	73	32	

(c) HAM10000

Data subclass	Train	Test	Total
0	228	66	294
1	359	103	462
2	769	220	989
3	80	23	103
4	779	223	1002
5	4693	1341	6034
6	99	29	128
Total	7007	2005	

Implementation Details

The experiments were conducted using the PyTorch framework and executed on an NVIDIA Tesla T4 GPU with 16 GB of RAM. The initial learning rate was set to 0.001, and the Adam optimizer updated the model parameters with a weight decay of 0.0005. All images in datasets were resized to 64 × 64. Additionally, random horizontal-flip or vertical-flip data augmentations were applied. All models were trained for 100 epochs for upstream tasks and 800 epochs for downstream tasks, using a batch size of 128.

The same settings were also used for the external validation of magnetic tile defect and HAM10000 dataset. For the HAM10000 with small input size was scaled to 32 × 32 to account for the minimum input size of 32 for typical anomaly models for effective learning and sufficient feature representation Ref. 33. The proposed and comparison models were trained for 100 epochs with batch size 2. To deal with the small size of input data, all the comparative models parameterized to the same network architecture with our proposed model, including latent vector size 64 and model depth 3.

Results

Anomaly Detection Performance Comparison

The anomaly performance of our proposed MT-former was compared with several state-of-the-art anomaly-detection models on the SK dataset (Table 3, Fig. 3). DAE achieved the lowest AUC (0.3220); this result indicates that DAE reconstruction scores do not work well on multi-normal classes because appropriate reconstruction is difficult when the normal shapes are diverse. The models that used GAN with discriminator show higher AUC (GANomaly 0.5897, AnoGAN 0.6912) than DAE. AnoGAN had higher TPR and AUC than GANomaly. Although GANomaly shows slightly lower FPR than AnoGAN, the lower TPR of GANomaly suggests that it mainly predicts the negative class. Patchcore demonstrates high TPR (0.7083) by leveraging a patch-level approach to analyze localized information. However, the FPR and AUC are poorer than our model, indicating challenges in handling multiple normal classes. A high FPR further reduces anomaly detection efficiency, especially in an imbalanced SK dataset, where the large number of normal cases leads to more misclassifications as abnormal. The base model, Deep-SVDD, exhibits a low TPR at the trained threshold, suggesting that the Deep-SVDD model is not well trained when various defect cases are considered as one class. Our proposed model, MT-former, shows the highest AUC (0.7821) by solving the problem of diverse defect types. Even if S/B metric of our proposed method is lower than GAN-based models, the proposed method achieved a higher AUC and TPR for overall small FPR than the other models (Fig. 3).

The Deep-SVDD based model (i.e., Deep-SVDD and MT-former) also has fewer parameters than models that use GAN (i.e., AnoGAN and Ganomaly) or based on patch-based models (i.e., Patch-SVDD and Patchcore). In terms of computational efficiency, MT-formers may be slightly slower than other models, but they offer high performance and compact model size with reasonably fast processing speed for automated defect detection. Since

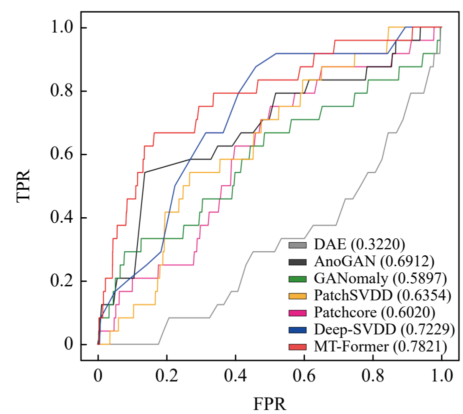

Fig. 3 Receiver operating characteristic (ROC) curve analysis of different state-of-the-arts and our proposed network on SK-defect dataset.

Table 3 Results of different state-of-the-arts and our proposed network on SK-defect dataset.

Model	SK-defect				Param #	Training Time (sec)	Inference Time (msec)
	TPR (↑)	FPR (↓)	AUC (↑)	S/B (↑)			
DAE[18]	0	**0**	0.3220	Inf	46K	20	5
AnoGAN[31]	<u>0.5417</u>	0.1369	0.6912	3.06%	167K	10	83
GANomaly[33]	0.2916	<u>0.0728</u>	0.5897	3.10%	910K	7	5
PatchSVDD[41]	<u>0.5417</u>	0.2684	0.6354	0.45%	106K	80	10
Patchcore[42]	**0.7083**	0.4402	0.6020	1.24%	68M	60	192
Deep-SVDD[34]	0	**0**	<u>0.7229</u>	Inf	23K	12	19
MT-former (Proposed)	**0.7083**	0.2491	**0.7821**	2.20%	66K	46	24

Notes. Training time shows the time required to train one epoch. Inference time indicates inference time per one image.

semiconductor manufacturing plants require significant resources, our lightweight MT-former model with high performance can be practically utilized for anomaly detection in this field.

In addition, the MT-former achieved the highest AUC (0.6798) on the magnetic tile defect dataset (Table 4 magnetic tile defect) and AUC (mean 0.7073) on the HAM10000 dataset (Table 4 HAM10000). Specifically, for the HAM10000 data, our model achieved the highest AUC scores in six of the seven classes except for class 5, demonstrating robust performance across different types of abnormal cases. These results show the model's generalization capabilities to a variety of external data sources.

The analysis of MT-former

In this section, we analyze the effectiveness of MTL, focal loss, and ESAM (Table 5). A basic Deep-SVDD model (Table 5a) had with TPR = 0 and FPR = 0; i.e., it identified all outcomes as 'normal'.

MTL applied upstream showed increased AUC (0.7751) with a slightly better TPR (0.1667) than basic Deep-SVDD (Table 5b) but did not adequately cluster features (Fig. 4b). In addition, an unstable validation error (Fig. 5b) still seems to limit the anomaly-detection accuracy. Compared to applying MTL upstream, applying MTL downstream increased TRP by 29% (Table 5c), and clustered each normal class (Fig. 4c). When MTL was applied both upstream and downstream (Table 5d), its FPR (0.1795) AUC (0.7268) and clustering of normal classes all improved (Fig.4d) with the stabilized error (Fig. 5c).

These results demonstrate that the basic Deep-SVDD does not detect new anomalies when multiple classes are present, whereas MTL that considers different normal class types increases overall detection accuracy. The violin plot of Deep-SVDD (Fig. 6a) shows that distance between the latent vectors and the center point is very short (3.8×10^{-7}) on y-axis compared to the other plots (Fig. 6b-d), and this distance does not even distinguish normal from abnormal at all. This result demonstrates that the Deep-SVDD only learns clustering to minimize the distance among all data without considering visual differences of data types. In this setting that includes a variety of normal classes, we found that separate clustering by types of normal classes in advance helps to increase detection of new types of anomalies.

We also make use of focal loss to compensate for data imbalance (Table 5e), and ESAM to extract global context

Table 4 AUC score of different state-of-the-arts and our proposed network on magnetic tile defect and HAM10000 external dataset.

Model	Magnetic tile defect	HAM10000							
		0	1	2	3	4	5	6	Mean
DAE[18]	0.3258	0.7220	0.6828	0.6214	0.4388	0.6376	0.3366	0.5807	0.5743
AnoGAN[31]	0.5360	0.5008	0.6279	0.4919	0.5769	0.4827	**0.5426**	0.5980	0.5444
GANomaly[33]	0.6644	0.6981	0.6505	0.6027	0.5582	0.5493	0.4042	0.4512	0.5592
PatchSVDD[41]	0.6357	0.4853	0.4842	0.5000	0.3908	0.6420	0.4707	0.4858	0.4941
Patchcore[42]	0.3840	0.4457	0.4004	0.3935	0.3758	0.4958	0.1646	0.3783	0.3791
Deep-SVDD[34]	0.3540	0.5103	0.5000	0.5444	0.5071	0.5073	0.4757	0.4502	0.4993
MT-former (Proposed)	**0.6798**	**0.7480**	**0.8193**	**0.6767**	**0.8020**	**0.6807**	0.4324	**0.7918**	**0.7073**

Table 5 Ablation studies of the proposed method on SK Hynix data.

	MTL-DAE	MTL-SVDD	Focal	ESAM	TPR (↑)	FPR (↓)	AUC (↑)
(a)					0	**0**	0.7229
(b)	√				0.1667	0.0342	0.7751
(c)		√			0.4583	0.3032	0.5489
(d)	√	√			0.4583	0.1795	0.7268
(e)	√	√	√		0.4583	0.2056	0.6658
(f)	√	√		√	0.3750	0.1579	0.6527
(g)	√	√	√	√	**0.7083**	0.2491	**0.7821**

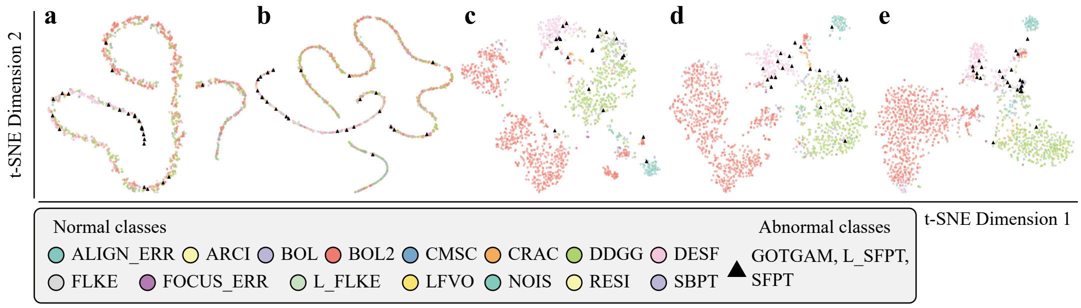

Fig. 4 Embedded feature visualization by t-SNE[65]. **a** Deep-SVDD. **b** Deep-SVDD with multi-task learning applied only to the upstream phase (DAE). **c** Deep-SVDD with multi-task learning to the downstream phase (SVDD). **d** Deep-SVDD with multi-task learning at both the upstream and the downstream phase. **e** Proposed MT-former with multi-task learning, focal loss, and ESAM.

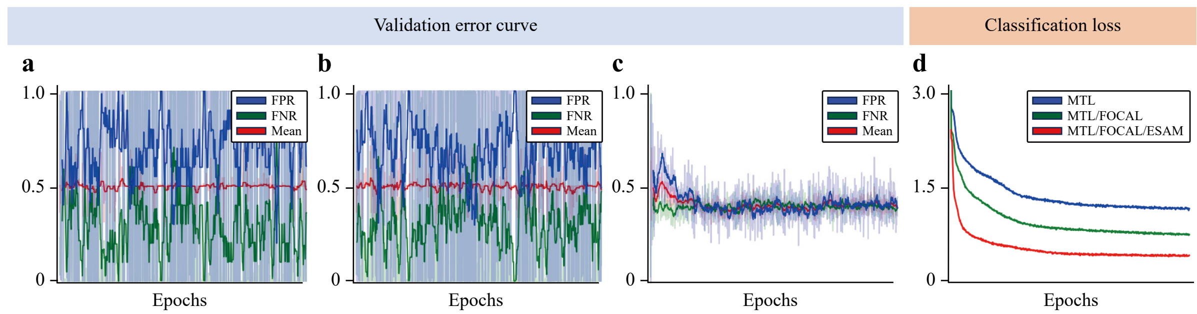

Fig. 5 Validation error curves for anomaly detection and classification task loss. **a** Deep-SVDD with ESAM. **b** Deep-SVDD with ESAM, including multi-task learning at the upstream phase. **c** Proposed MT-former with ESAM, including multi-task learning to both the upstream and the downstream phase. **d** Comparison of classification loss. a, b show fluctuation of FPR and FNR according to a training step.

(Table 5f). Compared with Table 5d, use of only focal loss did not increase TPR (Table 5e). Use of only ESAM slightly improves the FPR, but still limits the TPR (Table 5f). On the other hand, the combination of focal loss with ESAM achieved the highest TPR (0.7083) and AUC (0.7821) (Table 5g). From the results, it is illustrated that the simultaneous use of focal loss and ESAM module mutually leverage each method's effect by improving TPR. In other words, the ESAM module improves the efficiency of the focal loss. This effect can also be observed in (Fig. 4e), which shows that both normal and abnormal cases are better clustered than when MTL was used alone (Fig. 4d). Furthermore, embedded features show a significant difference between normal case and abnormal case ($p < 0.0001$) (Fig. 6d).

In addition, we conducted an ablation study that considered different numbers of heads for ESAM (Table 6). The best results (TPR = 0.7083, AUC = 0.7821) were obtained with three heads; i.e., both too many heads and too few heads impair ESAM results.

Normal class classification result

The effectiveness of MTL, focal loss, and ESAM on normal class classification accuracy was examined (Table 7). Overall accuracy was quantified using a weighted average of the AUCs of each class. As a baseline, a basic Deep-SVDD encoder was trained from scratch for classification purposes only (Table 7a). Fine tuning of an encoder extracted from MTL at both upstream and downstream (Table 7b) achieved the best score for the BOL class (0.8776), but poor results for most classes. The application of focal loss for imbalance multi-class data (Table 7c) achieved the highest score on L_FLKE (0.9487) and second best on five classes (ALIGN_ERR, ARCI, BOL2, FLKE, and NOIS), improved weighted average (0.8985) and showed better loss reduction (Fig. 5d).

Fine-tuning the encoder from MT-former achieved superior accuracy in most classes, and yielded the highest overall weighted average (0.9290) (Table 7d) and the best loss convergence (Fig. 5d). These results indicate that ESAM achieves the best identification of new defect types, and it improved its ability to classify irregular patterns by considering the global context. Furthermore, from the embedded features in the previous analysis (Fig. 4e), we deduce that effectively clustering each class with the MT-former can also be helpful in classification tasks.

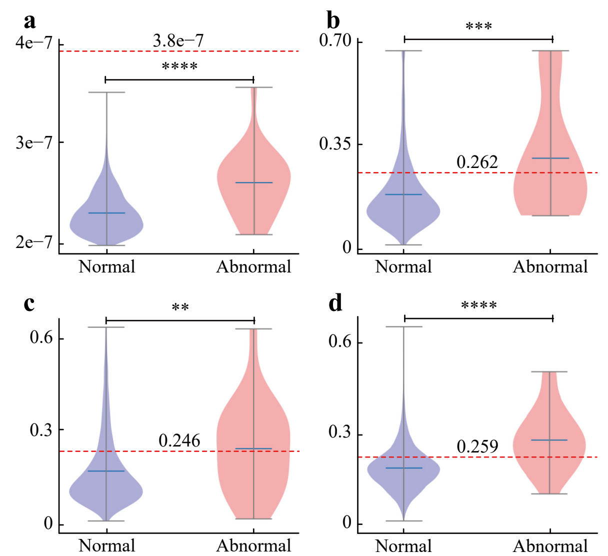

Fig. 6 Violin plots of distance distributions from a center point. The y-axis means distance between latent vectors of data and a center point. Red dotted line: threshold representing distance of decision boundary between abnormal (new types of defects) and normal (existing types of defects). **a** Deep-SVDD. **b** Deep-SVDD with multi-task learning applied both upstream and downstream. **c** Deep-SVDD with multi-task learning and focal loss. **d** Proposed MT-former with multi-task learning, focal loss, and ESAM. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.

Table 6 Ablation study on different number of attention heads on SK Hynix data.

Number of heads	TPR (↑)	FPR (↓)	AUC (↑)
1	0.4583	0.2555	0.6497
2	0.2917	**0.1985**	0.4696
3	**0.7083**	0.2491	**0.7821**
4	0.2917	0.2346	0.4993
5	0.4167	0.2775	0.5039

Conclusion

Conventional anomaly detection only distinguishes between non-defective and defective images, the models are just required to identify regular patterns of normal images. Deep autoencoders (DAEs), a classic anomaly detection model, perform poorly when applied to multi-normal classes, because it is difficult to accurately reconstruct different normal shapes. Deep Support Vector Domain Description (Deep-SVDD) is known to classify anomaly cases given single-normal case. Since the normal cases have similar shapes and patterns in conventional task, the latent vectors are clustered properly even if they are trained as one class. However, it is observed that the vectors are not properly clustered if the existing abnormal cases are heterogeneous in shape and pattern, as in our task. We analyzed that training diverse patterns of existing defects as a single class led to a broad data distribution on the existing defects, causing new defects to be included into the broad distribution. To solve this problem, we used multi-task learning (MTL) to simultaneously learn to distinguish the existing abnormal classes, avoiding the broad distribution and leading to the well-clustered latent vectors. In addition, since our model is only trained on the existing defects, the model does not require retraining when a new defect is identified. Considering intricate patterns of the current defect classes, we introduced MT-former that uses MTL and an ESAM to detect unknown defects that existing inspection systems cannot find in scanning electron microscope (SEM) images. MTL can simultaneously classify the existing defect kinds with various forms, so it can cluster existing classes efficiently for anomaly identification. MTL also greatly stabilizes training with respect to false positive rate (FPR) and false negative rate (FNR). ESAM takes global contextual features to consider irregular patterns from complex fabrication systems, and maximizes the efficiency of focal loss to effectively analyze imbalanced data.

Compared with SOTA models, our method shows better TPR result for especially the region < 20% FPR region with high AUC (Fig. 3), representing that our model

Table 7 Quantitative results of classification performance (AUC score) on existing defect cases.

Model	ALIGN _ERR	ARCI	BOL	BOL2	CMSC	CRAC	DDGG	DESF	FLKE	FOCUS _ERR	LFVO	L_FLKE	NOIS	RESI	SBPT	Weighted average
(a)	0.7544	0.9431	0.8684	0.7924	0.8338	0.8572	0.9299	**0.9700**	**0.9738**	**0.8287**	0.9774	0.9163	0.8408	0.8723	0.8128	0.8477
(b)	0.7896	0.9169	**0.8776**	0.9096	0.7674	0.8556	0.8510	0.9647	0.9116	0.7466	0.9485	0.8896	0.8412	0.7672	0.7200	0.8805
(c)	0.9108	0.9527	0.7868	0.9135	0.7955	0.7999	0.8980	0.9630	0.9582	0.7368	0.9708	**0.9487**	0.8986	0.6753	0.7808	0.8985
(d)	**0.9534**	**0.9627**	0.8664	**0.9357**	**0.8657**	**0.9306**	**0.9337**	0.9586	0.9188	0.8161	**0.9828**	0.8660	**0.9211**	**0.8987**	**0.8368**	**0.9290**

Notes. **a** Training an encoder of Deep-SVDD from scratch. **b** Fine tuning an encoder from multi-task learning at both the upstream and the downstream phase. **c** Fine tuning an encoder from multi-task learning with focal loss at both the upstream and the downstream stage. **d** Fine tuning an encoder from proposed MT-former with multi-task learning, focal loss, and ESAM.

provides more balanced performance at various thresholds. Table 3 demonstrates that while GANomaly records low FPR at an optimal threshold, the TPR remains below 50% (Table 3). This indicates that missing even a few defects could lead to significant economic losses in the semiconductor industry, making the model unsuitable for practical applications. Hence, our model proves to be more effective than other methods for this field. These improvements are attributed to the integration of the ESAM and convolutional module as a block, enabling effective extraction of both local and global features. As a result, our model successfully detects defects across both localized areas and broader regions, enhancing overall reliability. Finally, the pretrained model for anomaly detection demonstrates that the model can also serve as a weight-initialization technique to classify the existing-defect classes.

For future works, we discuss potential challenges. First of all, a small increase in FPR indicates a large number of normal cases are misclassified as abnormal because of the imbalanced data. Therefore, further improvements in reducing FPR are essential to achieve reliable anomaly detection. Second, the sub-sampling approach of the ESAM should be further studied. In its current implementation, the sub-sampling reduces resolution at a fixed ratio across all layers. However, applying the same subsampling rate to smaller feature maps of last layers can cause substantial loss of abstract information. Therefore, it is needed to study finding the optimal sub-sampling approach with less information loss. Last, the model's explainability should be covered more. While visualization techniques like t-SNE provide valuable insights into the model's behavior, incorporating methods such as Class Activation Mapping (CAM) could highlight the regions the model focuses on during classification or anomaly detection tasks. This advancement would improve interpretability and offer deeper insights into the model's decision-making process, fostering greater transparency and trust in its application.

Acknowledgements
This work was supported by SK Hynix AICC (P23.03); by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (2023R1A2C3004880) and the Ministry of Education (2020R1A6A1A03047902 and 2022R1A6A1A03052954); by Basic Science Research Program through the NRF funded by the Ministry of Education (RS-2024-00415450); by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)); by the BK21 FOUR project; by Glocal University 30 projects.

Author contributions
Hyunsu Jeong: Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. Chiho Yoon: Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. Hyunseok Lim: Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Writing – review & editing. Jaesuk Chang: Data acquisition, Writing – review & editing. Sampa Misra: Writing – original draft, Writing – review & editing. Chulhong Kim: Supervision, Writing – review & editing, Funding acquisition.

Data availability
All data are available from the corresponding authors upon reasonable request.

Conflict of interest
Chulhong Kim has financial interests in OPTICHO, which, however, did not support this work.

Received: 21 August 2024 Revised: 12 March 2025 Accepted: 25 March 2025
Accepted article preview online: 26 March 2025
Published online: 29 May 2025

References

1. Gómez-Sirvent, J. L. et al. Defect classification on semiconductor wafers using fisher vector and visual vocabularies coding. *Measurement* **202**, 111872 (2022).
2. Harada, M., Minekawa, Y. & Nakamae, K. Defect detection techniques robust to process variation in semiconductor inspection. *Measurement Science and Technology* **30**, 035402 (2019).
3. Bhonsle, R. et al. Inspection, characterization and classification of defects for improved CMP of III-V materials. *ECS Journal of Solid State Science and Technology* **4**, P5073-P5077 (2015).
4. Zipfel, J. et al. Anomaly detection for industrial quality assurance: a comparative evaluation of unsupervised deep learning models. *Computers & Industrial Engineering* **177**, 109045 (2023).
5. Chandola, V., Banerjee, A. & Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* **41**, 15 (2009).
6. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**, 21-27 (1967).
7. Carratù, M. et al. A novel methodology for unsupervised anomaly detection in industrial electrical systems. *IEEE Transactions on Instrumentation and Measurement* **72**, 3532812 (2023).
8. Krizhevsky, A. , Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA: Curran Associates Inc. , 2012, 1097-1105.
9. He, K. M. et al. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016, 770-778.
10. Yang, J. G. et al. Recent advances in deep-learning-enhanced photoacoustic imaging. *Advanced Photonics Nexus* **2**, 054001 (2023).
11. Park, J. et al. Clinical translation of photoacoustic imaging. *Nature Reviews Bioengineering* (2024) http://dx. doi.org/10.1038/s44222-024-00240-y.
12. Misra, S. et al. Deep learning-based multimodal fusion network for segmentation and classification of breast cancers using B-mode and elastography ultrasound images. *Bioengineering & Translational Medicine* **8**, e10480 (2023).
13. Yoon, C. et al. Collaborative multi-modal deep learning and radiomic features for classification of strokes within 6h. *Expert Systems with Applications* **228**, 120473 (2023).

14. Jeong, H. et al. Robust ensemble of two different multimodal approaches to segment 3D ischemic stroke segmentation using brain tumor representation among multiple center datasets. *Journal of Imaging Informatics in Medicine* **37**, 2375-2389 (2024).
15. Park, E. et al. Unsupervised inter-domain transformation for virtually stained high-resolution mid-infrared photoacoustic microscopy using explainable deep learning. *Nature Communications* **15**, 10892 (2024).
16. Kim, S. et al. Convolutional neural network–based metal and streak artifacts reduction in dental CT images with sparse-view sampling scheme. *Medical Physics* **49**, 6253-6277 (2022).
17. Misra, S. et al. Bi-modal transfer learning for classifying breast cancers via combined B-mode and ultrasound strain imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **69**, 222-232 (2022).
18. Choi, S. et al. Deep learning enhances multiparametric dynamic volumetric photoacoustic computed tomography in vivo (DL-PACT). *Advanced Science* **10**, 2202089 (2023).
19. Wang, M., Zhou, D. H. & Chen, M. Y. Hybrid variable monitoring mixture model for anomaly detection in industrial processes. *IEEE Transactions on Cybernetics* **54**, 319-331 (2024).
20. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504-507 (2006).
21. Bergmann, P. et al. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Funchal, Portugal: VISIGRAPP, 2019, 372-380.
22. Sakurada, M. & Yairi, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis. Gold Coast, Australia: ACM, 2014, 4-11.
23. Masci, J. et al. Stacked convolutional auto-encoders for hierarchical feature extraction. Proceedings of the 21st International Conference on Artificial Neural Networks and Machine Learning. Espoo, Finland: Springer, 2011, 52-59.
24. Zhang, H. B. et al. Unsupervised deep anomaly detection for medical images using an improved adversarial autoencoder. *Journal of Digital Imaging* **35**, 153-161 (2022).
25. Zhang, C. K. , Wang, Y. M. & Tan, W. M. MTHM: self-supervised multitask anomaly detection with hard example mining. *IEEE Transactions on Instrumentation and Measurement* **72**, 3518613 (2023).
26. Luo, J. X. et al. SMD anomaly detection: a self-supervised texture–structure anomaly detection framework. *IEEE Transactions on Instrumentation and Measurement* **71**, 5017611 (2022).
27. Cheng, X. et al. Deep self-representation learning framework for hyperspectral anomaly detection. *IEEE Transactions on Instrumentation and Measurement* **73**, 5002016 (2024).
28. Goodfellow, I. et al. Generative adversarial nets. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014, 2672-2680.
29. Kim, J. et al. Deep learning alignment of bidirectional raster scanning in high speed photoacoustic microscopy. *Scientific Reports* **12**, 16238 (2022).
30. Kim, G. et al. Integrated deep learning framework for accelerated optical coherence tomography angiography. *Scientific Reports* **12**, 1289 (2022).
31. Kim, J. et al. Deep learning acceleration of multiscale superresolution localization photoacoustic imaging. *Light*: *Science & Applications* **11**, 131 (2022).
32. Niu, M. H. et al. An adaptive pyramid graph and variation residual-based anomaly detection network for rail surface defects. *IEEE Transactions on Instrumentation and Measurement* **70**, 5020013 (2021).
33. Schlegl, T. et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. Proceedings of the 25th International Conference on Information Processing in Medical Imaging. Boone, NC, USA: Springer, 2017, 146-157.
34. Lee, S. et al. Emergency triage of brain computed tomography via anomaly detection with a deep generative model. *Nature Communications* **13**, 4251 (2022).
35. Akcay, S. , Atapour-Abarghouei, A. & Breckon, T. P. GANomaly: semi-supervised anomaly detection via adversarial training. Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer, 2019, 622-637.
36. Ruff, L. et al. Deep one-class classification. Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018, 4393-4402.
37. Misra, S. et al. A voting-based ensemble feature network for semiconductor wafer defect classification. *Scientific Reports* **12**, 16254 (2022).
38. Imoto, K. et al. A CNN-based transfer learning method for defect classification in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing* **32**, 455-459 (2019).
39. Chen, Z. Q. et al. DMVSVDD: multi-view data novelty detection with deep autoencoding support vector data description. *Expert Systems with Applications* **240**, 122443 (2024).
40. Dong, X. H., Taylor, C. J. & Cootes, T. F. Defect classification and detection using a multitask deep one-class CNN. *IEEE Transactions on Automation Science and Engineering* **19**, 1719-1730 (2022).
41. Liu, B. et al. Adaboost-based SVDD for anomaly detection with dictionary learning. *Expert Systems with Applications* **238**, 121770 (2024).
42. Zhou, Y. et al. VAE-based deep SVDD for anomaly detection. *Neurocomputing* **453**, 131-140 (2021).
43. Yi, J. H. & Yoon, S. Patch SVDD: patch-level SVDD for anomaly detection and segmentation. Proceedings of the 15th Asian Conference on Computer Vision. Kyoto, Japan: Springer, 2020, 375-390.
44. Roth, K. et al. Towards total recall in industrial anomaly detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE, 2022, 14298-14308.
45. Yu, J. B. & Liu, J. T. Two-dimensional principal component analysis-based convolutional autoencoder for wafer map defect detection. *IEEE Transactions on Industrial Electronics* **68**, 8789-8797 (2021).
46. Kang, H. & Kang, S. A stacking ensemble classifier with handcrafted and convolutional features for wafer map pattern classification. *Computers in Industry* **129**, 103450 (2021).
47. Cheon, S. et al. Convolutional neural network for wafer surface defect classification and the detection of unknown defect class. *IEEE Transactions on Semiconductor Manufacturing* **32**, 163-170 (2019).
48. Wen, G. J. et al. A novel method based on deep convolutional neural networks for wafer semiconductor surface defect inspection. *IEEE Transactions on Instrumentation and Measurement* **69**, 9668-9680 (2020).
49. Kim, E. S. et al. An oversampling method for wafer map defect pattern classification considering small and imbalanced data. *Computers & Industrial Engineering* **162**, 107767 (2021).
50. Tao, X. et al. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement* **71**, 5018021 (2022).
51. Gao, Y. P. et al. A multilevel information fusion-based deep learning method for vision-based defect recognition. *IEEE Transactions on Instrumentation and Measurement* **69**, 3980-3991 (2020).
52. Yang, L. M., Zhou, F. Q. & Wang, L. A scratch detection method based on deep learning and image segmentation. *IEEE Transactions on Instrumentation and Measurement* **71**, 5015012 (2022).

53. Tao, X. et al. ViTALnet: anomaly on industrial textured surfaces with hybrid transformer. *IEEE Transactions on Instrumentation and Measurement* **72**, 5009013 (2023).
54. Shang, H. B. et al. Defect-aware transformer network for intelligent visual surface defect detection. *Advanced Engineering Informatics* **55**, 101882 (2023).
55. Vaswani, A. et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc. , 2017, 6000-6010.
56. Gao, Y. H. , Zhou, M. & Metaxas, D. N. UTNet: a hybrid transformer architecture for medical image segmentation. Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg, France: Springer, 2021, 61-71.
57. Wang, S. N. et al. Linformer: self-attention with linear complexity. Print at https://arxiv.org/abs/2006.04768 (2020).
58. Bello, I. et al. Attention augmented convolutional networks. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019, 3285-3294.
59. Shaw, P. , Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, LA, USA: ACL, 2018, 464-468.
60. Caruana, R. Multitask learning. *Machine Learning* **28**, 41-75 (1997).
61. Lin, T. Y. et al. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017, 2999-3007.
62. Dureuil, V. et al. Wafer bevel shape inducing high defect density in shallow trench isolation process. Proceedings of 2010 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC). San Francisco, CA, USA: IEEE, 2010, 213-216.
63. Huang, Y. B., Qiu, C. Y. & Yuan, K. Surface defect saliency of magnetic tile. *The Visual Computer* **36**, 85-96 (2020).
64. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**, 180161 (2018).
65. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).