Deep Learning for Unsupervised Anomaly Localization in Industrial Images: A Survey

Xian Tao, *Member, IEEE*, Xinyi Gong, Xin Zhang, Shaohua Yan, and Chandranath Adak, *Senior Member, IEEE*

***Abstract*— Currently, deep learning-based visual inspection has been highly successful with the help of supervised learning methods. However, in real industrial scenarios, the scarcity of defect samples, the cost of annotation, and the lack of *a priori* knowledge of defects may render supervised-based methods ineffective. In recent years, unsupervised anomaly localization (AL) algorithms have become more widely used in industrial inspection tasks. This article aims to help researchers in this field by comprehensively surveying recent achievements in unsupervised AL in industrial images using deep learning. The survey reviews more than 120 significant publications covering different aspects of AL, mainly covering various concepts, challenges, taxonomies, benchmark datasets, and quantitative performance comparisons of the methods reviewed. In reviewing the achievements to date, this article provides detailed predictions and analysis of several future research directions. This review provides detailed technical information for researchers interested in industrial AL and who wish to apply it to the localization of anomalies in other fields.**

***Index Terms*— Anomaly localization (AL), deep learning, industrial inspection, literature survey, unsupervised learning.**

I. Introduction

AUTOMATED visual inspection based on deep learning technology is being widely applied in industrial defect detection applications due to its efficiency and remarkable accuracy, including unmanned aerial vehicle (UAV) patrol inspection of power equipment [1], weak scratches detection on industrial surfaces [2], identification of copper wire defect in deep hole parts [3], conductive particle detection for chip on glass [4], and so on. Existing inspection systems are primarily based on the supervised learning method, which significantly relies on labeled data. Image category labels, bounding box labels, and fine-grain pixelwise labels are the three classical types of labels available. Unfortunately, the abovementioned fully supervised approaches suffer from several inevitable limitations.

1) The ample annotations are labor-intensive and high cost.

2) As the process on several precision generation lines improves, defective samples are becoming scarce, posing labeling challenges.

3) All possible defective types need to be known in advance under fully supervised learning.

4) Annotation noise may be inadvertently introduced when labeling the data.

As a result, both academia and industry have paid extensive attention to develop unsupervised technology for vision inspection systems.

A. Anomaly Detection Versus Anomaly Localization

The human visual system has the inherent ability to perceive anomalies—not only can humans distinguish between defective and nondefective images, even if they have never seen any defective samples before, but they can also point out the location of anomalies. Anomaly localization (AL) was introduced to academia for the very same purpose, i.e., to teach the machine to "find" the anomaly region in an unsupervised manner. In the context of deep learning methods, "unsupervised" means that the training stage contains only normal images without any defective samples. AL method under the unsupervised paradigm first avoids the hardship of collecting anomalous or defective samples, which cannot be avoided in the supervised method, since the normal images without defects are far more than the abnormal samples in the industrial scenario. Second, the labeling cost of the training sample in the supervised method can be eliminated in the unsupervised method. Last but not least, the unsupervised method also avoids the influence of labeling deviation, which is commonly seen in the supervised method. Since the training data only have the normal class, it may be called "semi-supervised." However, to unify with most of the current methods, we remove the term "unsupervised" or "semi-supervised" in the following content and only call it AL. The distinction between AD and AL is depicted in Fig. 1. Outlier detection or one-class classification is the other term for AD. It refers to the task of distinguishing defective images at the image level from the majority of nondefective images. AL, on the other

Manuscript received 25 February 2022; revised 1 July 2022; accepted 20 July 2022. Date of publication 4 August 2022; date of current version 16 August 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103004, in part by the Beijing Municipal Natural Science Foundation (China) under Grant 4212044, and in part by the National Natural Science Foundation of China under Grant 62066004. The Associate Editor coordinating the review process was Dr. Zhibin Zhao. *(Corresponding authors: Xian Tao; Chandranath Adak.)*

Xian Tao, Xinyi Gong, and Shaohua Yan are with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: taoxian2013@ia.ac.cn).

Xin Zhang is with the Key Laboratory of Industrial Internet and Big Data, China National Light Industry, Beijing Technology and Business University, Beijing 100048, China.

Chandranath Adak is with the Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihar 801106, India (e-mail: chandranath@iitp.ac.in).

Digital Object Identifier 10.1109/TIM.2022.3196436

TABLE I
SUMMARY OF PREVIOUS REVIEWS

Title	Year	Venue	Description
Image Anomalies: A Review and Synthesis of Detection Methods [6]	2019	JMIV	This paper reviews the classical image AD models before 2018 and compares 6 representative algorithms on a synthetic database.
Deep learning for anomaly detection: A survey [7]	2019	Arxiv	This paper reviews deep learning-based AD methods along with their application across various domains.
A Unifying Review of Deep and Shallow Anomaly Detection [5]	2020	Proc. IEEE	This paper established a systematic unifying view of deep and shallow AD models and discussed many practical aspects.
Image/Video Deep Anomaly Detection: A Survey [8]	2021	Arxiv	This paper conducts an in-depth investigation into the images/videos of deep learning-based AD methods and discusses challenges and future research directions.
Deep learning for anomaly detection: A review [9]	2021	ACMCS	This paper surveys deep AD with a comprehensive taxonomy, covering advancements in 3 categories and 11 fine-grained categories of the methods.
Visual Anomaly Detection for Images: A Systematic Survey [10]	2022	Procedia CS	This paper provides a short survey of the classical and deep learning-based approaches for visual AD and AL.
GAN-based Anomaly Detection: A Review [11]	2022	Neurocomputing	This paper focus on GAN-based AD and discusses its theoretical basis and applications.

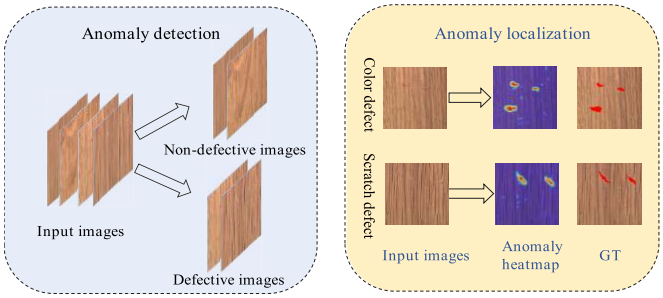

Fig. 1. Anomaly localization versus anomaly detection (the samples are from the wood dataset of MVTec AD [115]).

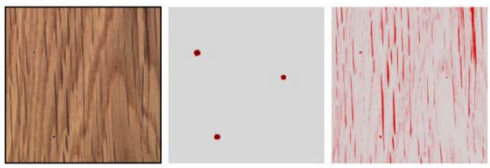

Fig. 2. Apparent limitations of anomaly detection. (a) Input image. (b) Ground truth of anomaly (three drill holes). (c) Visual explanation of the anomaly prediction. Model assigns high relevance to the normal texture of wood instead of the drill holes.

hand, is also known as anomaly segmentation, which is used to produce pixel-level anomaly location results. The darker the color in the anomaly heat-map, as shown in Fig. 1, the more likely the location is to be anomalous.

The AD task is insufficient to ensure that the method can identify the actual defect locations in real-world industrial scenarios. As AD only performs a binary classification of the image, the results are uninterpretable. Although the image is classified as a defect catalog, the focus areas of the network may not be abnormal. As illustrated in Fig. 2, the anomaly detection method tends to place a high value on wood strains rather than on drill holes that are the real anomalies. Finding anomalies on industrial scene images is the starting point for this survey.

B. Differences From Previous Surveys

Table I enlists the existing surveys that are similar to ours. Ehret *et al.* [6] reviewed classical approaches up to 2018 and thus did not include recent deep learning-based solutions. Chalapathy and Chawla [7] investigated deep AD in supervised, semi-supervised, and unsupervised domains. Ruff *et al.* [5] provide a recent comprehensive review of the connections between traditional "shallow" and new "deep" approaches to AD. Mohammadi *et al.* [8] provided an overview and classification of image/video deep learning-based AD methods, dividing them into three categories: self-supervised learning (SSL), generative networks, and anomaly generation. Pang *et al.* [9] provided a comprehensive taxonomy for deep AD, covering advances in methods for three categories and 11 refinement categories. Yang *et al.* [10] presented a concise overview of traditional and deep learning-based visual AD and AL techniques. Xia *et al.* [11] conducted comprehensive survey research on generative adversarial networks (GANs)-based AD.

Multiple surveys related to AD/AL are presented in Table I, involving studies in the areas of early nondeep learning AD methods [6], deep crude AD methods [5], [7], [8], [9], limited AL models [10], or focusing only on GAN [11]. However, few surveys are dedicated to comprehensive AL methods. On the other hand, most of the existing surveys cast existing approaches into AD methods for image-level classification. As shown in Fig. 2, major AD methods easily ignore abnormal regions in industrial scenarios. Moreover, in recent five years, AL methods have developed from the image-level comparison (reconstruction or generation) to feature-level comparison and also from the simple proxy task of defect synthesis to the self-supervised method based on contrast learning. While AL on images or videos has been widely concerned, to the best of our knowledge; still no published paper has summarized detailed improvement and trends, e.g., network structure, loss function, feature comparison method, sample synthesis approach, and so on. Our work systematically and comprehensively reviews recent advances in unsupervised AL. It includes an in-depth analysis and discussion of numerous aspects that have never been explored in this area before, to the best of our knowledge. In particular, we summarize and discuss the existing methods for deep AL tackling various problems and challenges, provide a road map and taxonomy, review the existing datasets and evaluation metrics, present a

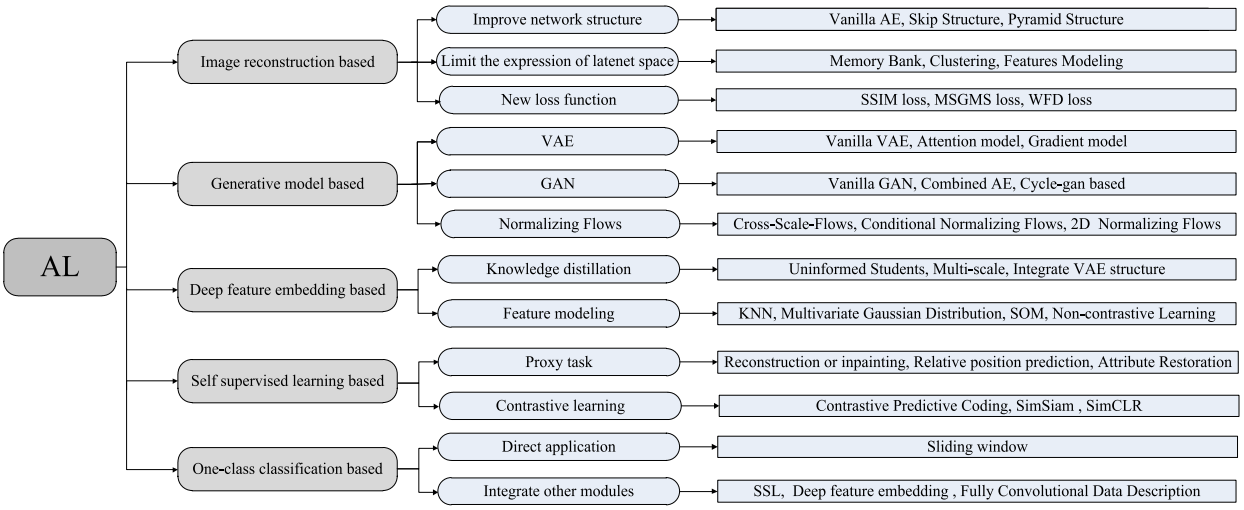

Fig. 3. Taxonomy of deep learning-based methods for unsupervised AL.

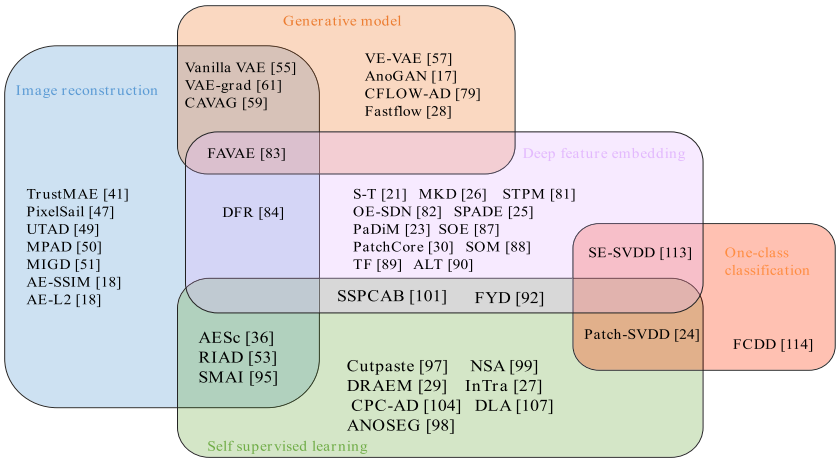

Fig. 4. Venn diagram of past major AL methods, divided into five categories. Some methods fall into more than one category, listed in the overlapping region.

comprehensive performance comparison of the state-of-the-art methods, and offer insights into future directions. We expect our survey to provide novel insight and inspiration, facilitating knowledge of deep AL and encouraging study on the open topics presented here.

C. Contributions to This Article

This article summarizes the surprising success and dominance of deep AL in industrial images but excludes other areas, such as medical images [12] and video AL [13]. Although some of the strategies have been validated in the above scenario, real industrial images lack *a priori* knowledge of medical images and video sequence information. The following are the main contributions of this article.

1) To the best of our knowledge, this is the first work to focus specifically on deep learning algorithms for unsupervised AL using industrial images. At present, most of the surveys on AL or surface defect detection in industrial scenes focus on supervision methods.

2) We present a taxonomy (refer to Fig. 3) that covers the latest and most advanced methods in deep learning for AL. We give a more detailed subclassification framework than previously outlined. In addition, we draw the typical AL network under the unsupervised paradigm with a Venn diagram (refer to Fig. 4), which is convenient for readers to understand the distinction and correlation between methods.

3) A comprehensive comparison of the existing methods on a public dataset is provided, while we also present a summary and insightful discussion.

The rest of this article is structured as follows. Section II summarizes the problems and related developments in AL during the previous five years. After that, Section III contains a taxonomy of the existing deep learning-based approaches. The following Section IV summarizes the benchmark datasets and presents an overview of evaluation metrics and their corresponding performance. Finally, Section V concludes this article with an important future research outlook.

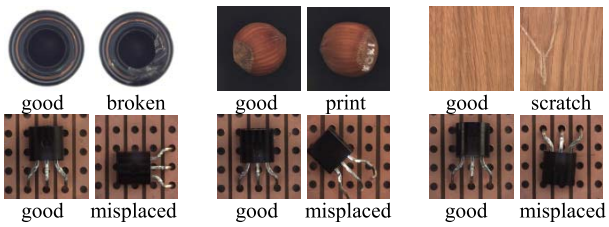

Fig. 5. Illustration of the types of anomalies: textural anomalies and functional anomalies. Low-level textural anomaly: a scratch on the wooden surface. High-level semantic anomaly: transistor with an offset pose or a pin not inserted in the hole. Note that the texture of a misplaced transistor is similar to that of a normal/good sample.

II. Background

A. Problems and Challenges

The goal of AL is to find anomaly regions using only defect-free training samples. The anomalies are defined as the observations that deviate significantly from some notion of normalcy. In general, anomalies in industrial scenarios are divided into two types: 1) textual anomalies with little semantic information and 2) functional anomalies with a considerable amount of semantic information. To better illustrate this distinction, we use images from the MVTec AD [20]. Texture anomalies account for a large percentage of industrial defect detection, such as cracks in bottles, marks on hazelnuts, and scratches on the surface of the wood. These can be regarded as variants of local pixels on the overall texture, as shown in the first row of Fig. 5. Functional anomalies differ from textural anomalies, which often do not have textural variations but contain semantic information. In the second row of Fig. 5, a subtle anomaly is concerned with whether the needle is inserted into the hole. This anomaly necessitates high-level semantic information, making it more difficult to detect than a textural anomaly.

In the literature, multiple closer terminologies are used, such as image segmentation, image saliency detection, surface defect defection, and novelty detection. We here explain the difference between AL with the other terminologies. Image segmentation is a broad concept. To some extent, AL is equivalent to unsupervised image segmentation, but image segmentation mostly focuses on acquiring specific objects with semantic information, which may not be abnormal. Image saliency detection can be defined as the task of finding saliency regions, which often correspond to the important object in the image. However, some anomalies may not be salient in the whole image, such as the functional anomalies in Fig. 5. Surface defect detection and anomaly location are very close concepts in industrial scenes. We can simply regard the anomaly location of industrial images as quite equivalent to unsupervised pixel-level defect detection. Novelty detection refers to image-level classification settings, where the inlier and outlier distributions differ significantly, which is quite similar to AD.

Moreover, there are significant challenges for AL in real industrial scenarios as follows.

1) Training Sample Distribution Problem: All the training samples used for unsupervised AL are defect-free. The degree of balance in the distribution of defect-free samples influences the judgment of anomaly location; for example, if a particular normal sample or region is missing from the training data, the trained model may identify that normal sample or region as anomalous. In other words, the goal is to make the machine's perspective as compatible with human experience as possible. Furthermore, there is the possibility of contamination or data noise in normal data in complex industrial scenarios. Variations in imaging conditions, such as illumination, perspective, scale, shadows, blur, and so on, can result in significant differences in training samples that should not be considered anomalies.

2) Multiscale Anomaly Problem: In real industrial scenes, some anomalies, such as cracks, are often subtle and occupy a tiny area. These small areas may even occupy only a few pixels in the entire high-resolution image. Thus, in anomalous images, tiny pixels are easily overwhelmed by normal conditions rather than anomalies. Furthermore, large span anomalies are also common in real-world scenes. It is, therefore, a challenge to locate anomalies by taking into account both small, subtle defects and large defects with a complete span.

3) Fine Boundary Problem: The decision boundary of the model should be equal to the ideal distribution boundary. However, because of the scarcity of pixelwise supervised labels, comprehensive segmentation of precise anomaly contours is another challenge in anomaly location. Currently, most anomaly location methods' location accuracy is insufficient, significantly different from the ground truth.

B. Road Map of Anomaly Localization

AL for industrial images has a brief history, dating back to the research of [14], [15], [16]. Most nondeep-learning-based AL models rely on sparse coding [14], [15] and dictionary learning [16]. Since 2017, a growing number of deep AL methods [19] have emerged due to the rousing success of deep learning techniques in computer vision. GAN models [17], [22] and AE reconstruction networks [18] were first used in the deep AL models. To consistently compare the effects of AL, a complete industrial AL dataset was proposed by MTVec company [20]. Later, feature embedding-based models, which are more effective and efficient, became the prevalent AL architecture. Knowledge distillation [21], [26] and pretrained feature comparison [23], [25], [30] are the examples of representative models. Then, several SSL-based methods have been applied to AL tasks [24], [29]. Flow-based models [28] and transformer models [27] as better approaches have also been embedded into AL networks. A brief chronology of AL is shown in Fig. 6. Despite its short history, AL research has produced hundreds of papers, and we have comprehensively selected influential papers published in prestigious journals and conferences; this survey focuses on major advances in the past five years.

III. Taxonomy

This section summarizes the unsupervised AL methods in terms of the high-level paradigm. Specifically, we review the various types of AL models given in Fig. 3, with subsections devoted to each category. In each subsection, we take

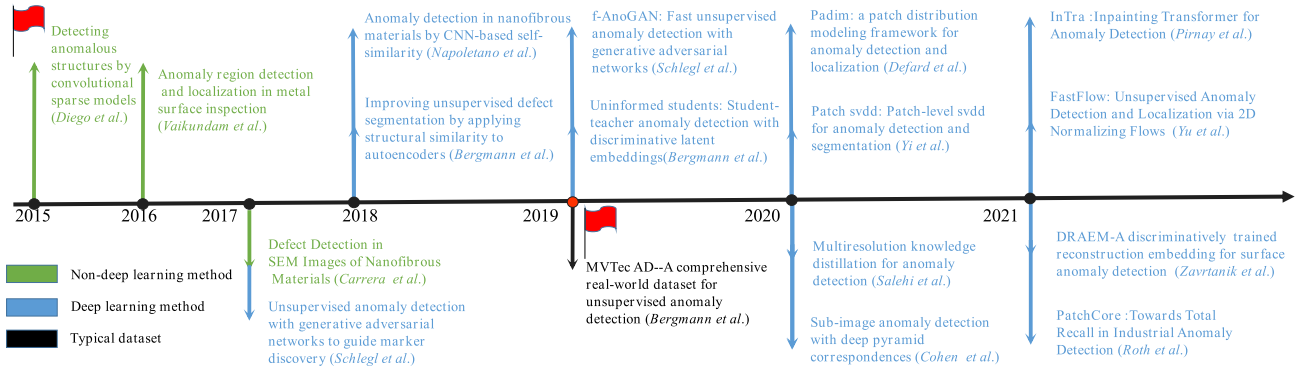

Fig. 6. Brief chronology of AL. Typical highly cited milestone methods are mentioned.

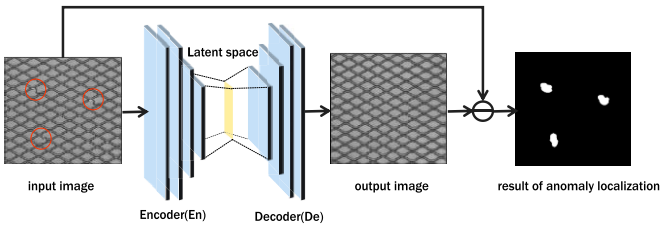

Fig. 7. Illustration of the main components of an AE.

a further breakdown of its representative works. However, some of the work fall into more than one category. Therefore, in Table X, we divide the works according to the Venn diagram of Fig. 4, and the overlapping area includes the cross part of the methods.

A. Image Reconstruction-Based Approach

The first group is the "image reconstruction-based methods," which are the most basic AL methods. It is based on the idea that the model is trained to reconstruct only normal images; and then, when an abnormal image is input, the model still reconstructs the anomalous region as normal, i.e., the model cannot reconstruct the abnormal image correctly. Therefore, the difference between the input and reconstructed images represents the localization result. As illustrated in Fig. 7, the input image is compressed on a low-dimensional bottleneck layer (latent space). This model assumes that the data have a high degree of correlation/structure. Consequently, the encoder compresses the data into an intermediate representation, which is then employed by the decoder to reconstruct the input image.

The earliest vanilla AE was used for unsupervised anomaly segmentation with brain MR images [31]. The industrial image reconstruction-based methods in AL follow this idea of the AE series. Youkachen *et al.* [32] used a convolutional autoencoder (CAE) for industrial image reconstruction. By sharpening the differences between the reconstructed and input images, they generated the final segmentation results of the surface defects in the hot-rolled strip. Kang *et al.* [33] reconstructed overlapping patches instead of the insulator image for detecting insulator surface defects, since the direct reconstruction of the entire image is intractable, and the defective area is usually a tiny part. Chow *et al.* [34] presented an application of deep learning in implementing AL of structural concrete defects to facilitate visual inspection of civil infrastructure. It also crops the input image and then feeds patches to vanilla AE for reconstruction. However, these vanilla AE methods may suffer from challenges due to the complex industrial scenarios. Here, we summarize the novel designs of the AE-based image reconstruction framework for AL.

1) Improvement of Network Structure: Different from the vanilla AE, two simple structure improvements are proposed to enhance the reconstruction ability better. The first one is skipping layers. Skip-GANomaly [35] employed an encoder–decoder convolutional neural network (CNN) with skip connections to thoroughly capture the multiscale distribution of the normal data distribution in high-dimensional image space. Based on an evaluation across multiple datasets from different domains and complexity, the skip connections provide more stable training and achieve numerically superior results than vanilla AE. Collin and de Vleeschouwer [36] proposed an AE architecture with skip connections for AD in industrial vision to increase the sharpness of the reconstruction. Besides, some works extended the design of AE and feature pyramid combination to the multiscale anomaly perception. Mei *et al.* [37] reconstructed image patches at different Gaussian pyramid levels with AE and synthesized the reconstructed results from these different resolution channels. Yang *et al.* [38] proposed a multiscale feature clustering-based fully CAE (MS-FCAE) method, which utilizes multiple feature AE subnetworks at different scale levels to reconstruct several textured background images. Mishra *et al.* [39] focused on image AD using a deep neural network with multiple pyramid levels to fuse image features at different scales.

However, the aforementioned improving structures do not work well on complicated textured or object datasets. Since some research shows that as AE employs a bottleneck layer to reconstruct the input image, it is difficult to manage its generalization ability. When the AE's generalization ability is powerful, anomalous features are confused with the normal

Fig. 8. Illustration of the strong generalization of AE.

feature, resulting in the network's output exactly reproducing the input. As depicted in Fig. 8, these models tend to directly copy the scratch area (marked with a red rectangle) as the output, resulting in missing anomalies. Due to the above reasons, many contemporary methods attempt to constrain latent space representation.

2) Constraining the Representation of Latent Space: According to how to deal with the features, we further group these methods into memory banks, clustering, and features modeling.

1) Memory banks employ a form of dictionary learning to replace the original expression of the latent space. Gong *et al.* [40] were the first to employ memory banks to detect anomalies. The memory bank module in this model is a matrix with each element similar to a word in dictionary learning and capable of encoding defect-free sample features. In particular, only a limited number of words are used for reconstruction during the training phase, prompting each matrix element to represent each row. Hence, the normal samples are indexed to the most comparable elements for good reconstruction, while the difference between the abnormal sample and reconstruction is magnified as the anomaly score. Later on, many following works [41], [42], [43], [44] adopted this design. Unlike prior memory bank approaches, SAP2 [45] constructed memory banks from pretrained features for AD and localization. Liao *et al.* [46] proposed a new AL framework by learning latent representations with selecting and weighting in a batch operation. This model is essentially a simplified version of a memory bank.

2) Clustering of latent space features is another way to enhance the discrimination of the model. Yang *et al.* [38] proposed a feature clustering module in MS-FCAE to enhance the discriminability of the encoded features in the latent space, which improves the reconstruction accuracy of the texture background image. An anomaly feature-editing-based adversarial network for texture defect visual inspection is proposed in [48], in which the latent space of the AE module also utilized feature clustering. Moreover, some classical clustering operations for latent space have been proposed, including the standard K-means clustering [50].

3) Modeling features of the latent space is also an effective way of limiting the representation. In [47], a discrete latent space probability model was estimated using a deep autoregressive model, named PixelSail. It determines the latent input space regions that deviate from the normal distribution during the detection stage. In particular, the deviation code is then resampled from the normal distribution and decoded to provide a restored image closest to the anomalous input. The anomaly region is identified by comparing the restored and anomaly images. In addition, some approaches to modeling latent space features have been proposed, including the Gaussian descriptors [51] and even graph network models [52].

As the image reconstruction-based methods usually employ a pixel-level comparison metric, AE networks choose trained loss with L_1-distance and L_2-distance. This results in the comparison of inputs and outputs being only at the pixel level and lacking semantic information. Therefore, some improvements based on the loss function have been proposed. For this kind of method, the key problem is considering the semantic information in the image reconstruction effect. We discuss this problem in the following.

3) New Loss Function: Bergmann *et al.* [18] were the first to use the structural similarity (SSIM) metric in image reconstruction. In contrast to pixelwise comparisons, SSIM loss considers a region's brightness, contrast, and structural information. Compared to L_2 loss, the SSIM loss significantly improves the performance of AL in the textured datasets. In [53], a new multiscale gradient magnitude similarity (MSGMS) loss was proposed, which pays more attention to the structural differences in the reconstruction. The MSGMS loss is constructed by calculating the gradient images of the original and reconstructed images. The overall area under the receiver operating characteristic curve (AUROC) on the MVTec AD is improved by 6.5% when using MSGMS. Nakanishi *et al.* [54] designed a new loss function named weighted frequency domain (WFD) loss, which transforms the reconstruction loss calculation from the image domain into the frequency domain. It provides a sharper reconstructed image, improving anomaly location accuracy.

Brief Summary: Table II gives a glance at these three types of image reconstruction-based methods and analyzes their advantages and disadvantages. Although image reconstruction-based methods are usually very intuitive and interpretable, their performance is limited because AE does not introduce any prior knowledge, and its effect only depends on the expression ability of latent layer to defect-free features.

B. Generative Model-Based Approach

In order to overcome the shortcomings of AE-based methods with poor reconstruction performance, generative models are introduced into the industrial AL field. The basic idea behind generative models is to model the real data distribution from the training data and then utilize the learned model and distribution to generate or model new data. The key to AL in this framework is explicitly or implicitly obtaining the feature distribution of defect-free data. As the generative model only generates normal samples, the difference between the generated or reconstructed samples and the input is the abnormal region. Unlike AE, which only considers the final reconstruction, the generative model could reflect this difference in latent or feature space. According to different models, we further group these methods into variational AE (VAE), GAN, and normalizing flow (NF).

1) Variational Autoencoder (VAE): As depicted in Fig. 9, VAE introduces a prior distribution for normal samples in the latent space, which is typically a multidimensional standard

TABLE II
STRENGTHS AND WEAKNESSES OF DIFFERENT IMAGE RECONSTRUCTION-BASED AL APPROACHES

Taxonomy	Methods	Strengths	Weaknesses
Improvement of network structure	Skip layers	Using a skip layer to enhance reconstruction performance	Reconstruction easily fails in complicated textured or object datasets.
	Feature pyramid	Suitable for multi-scale anomalous regions	
Constraining the representation of latent space	Memory banks	Reduces the strong generalization ability of the AE network	Hard to determine the optimal constrain parameters and is limited by the effect of pixel-level comparison.
	Clustering	Aims to cluster the effective information for latent representation	
	Modeling features	Restrict the distribution of latent space to a specific distribution	
New loss function	SSIM [18]	Add the luminance, contrast, or structure information for loss	Localization effect has not been significantly improved compared with the original loss.
	MSGMS [53]	Introducing the image gradient information for loss	
	WFD [54]	Transfer the image to the frequency domain to calculate the loss	

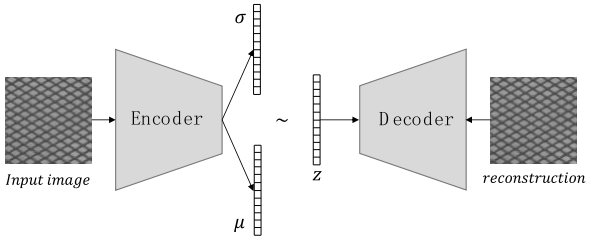

Fig. 9. Graphical illustration of VAE.

normal distribution. This indicates that the encoder output is no longer simply latent space but rather an estimated distribution, and therefore, this approach is a subset of modeling the latent space features in the AE-based methods. Thus, the difference between vanilla VAE and AE is the additional loss employed to assess the difference between the estimated distribution and the prior distribution, such as Kullback–Leibler (KL) divergence loss. Matsubara *et al.* [55] first introduced VAE for industrial AL, which was evaluated on a toy dataset and real-world manufacturing datasets. Kozamernik *et al.* [56] proposed a VAE-based model for visual quality control of electric cathode metal coating (KTL). By calculating the negative log-likelihood of the distribution returned by the decoder, anomalies containing surface defects are successfully detected. Although the vanilla VAE successfully located the anomalies, the localized accuracy for anomaly regions in [55] and [56] is relatively poor. Some researchers have tried to add other mechanisms to VAE for more fine-grained AL.

a) Attention-based methods: Liu *et al.* [57] first proposed a technique to generate VAE visual attention using gradient-based attention computation. The generation method of attention map is similar to grad-CAM [58]. In particular, the corresponding weight coefficient is obtained based on computing gradients of the latent space variable with respect to the last layer feature maps of the encoder. The final attention map is then generated by weighting the feature map of the last layer of the encoder. The apparent region in the acquired attention map is the anomalous one when detecting anomalous images. Venkataramanan *et al.* [59] proposed a convolutional adversarial VAE with guided attention (CAVGA), which localizes the anomaly with a latent convolutional variable to preserve the spatial information. It generates an attention map following the main idea of [57], with the expectation that the attention map generated by the training network could cover the entire image.

b) Gradient-based methods: According to Zimmerer *et al.* [60], the loss gradient with respect to input image gives the direction toward normal data samples, and its magnitude could indicate how abnormal a sample is. Benefiting from this conduction, Dehaene *et al.* [61] proposed the gradient descent-based VAE. As seen from the reconstructed images in [61], the gradient descent-based method gives better quality reconstructions than vanilla VAE. In [123], Chu and Kitani proposed that the change in loss values during training can also be used as a feature to identify anomalous data. The algorithm is thoroughly evaluated and compared against other baselines on two datasets, MVTec AD [115] and NanoTWICE [16], which spans a large variety of different objects and textures.

2) Generative Adversarial Network (GAN): GAN-based models are classified into three types based on their network structure as follows.

a) Vanilla GAN: Schlegl *et al.* [17] were the first to apply GAN to localize anomalies. The generative network G in this approach receives randomly sampled samples from the latent space as an input, and its output must be as close to the real samples in the training set as possible. The discriminative network D takes input from either the real samples or the output of the generative network, and its goal is to differentiate the output of the generative network from the real samples as much as possible. The whole loss includes two parts: the reconstruction loss of the G and the feature difference loss of the D. The difference between the output of the generating network G and the input image determines anomalous regions. Later on, several following works [62], [63] adopted this model for the industrial surface defect.

b) GAN combined with AE: As the vanilla GAN employs a single image as an input in the inference stage, the network must frequently repeat to find the optimal latent space vector to achieve the desired generation result. Some joint AE structured GAN methods have been proposed in response to the drawback that vanilla GAN needs to update its parameters constantly.

1) Improving the input of generator G is the most straightforward way to train a GAN-based AL network, where the input is changed to a real defect-free image instead of the randomly sampled samples from the latent space, and therefore, the generative network G is accordingly changed to a complete encoding–decoding structure, as shown in Fig. 10(A1). This improvement is equivalent to employ a discriminator D on the image reconstruction method to distinguish whether the image is a real input defect-free sample

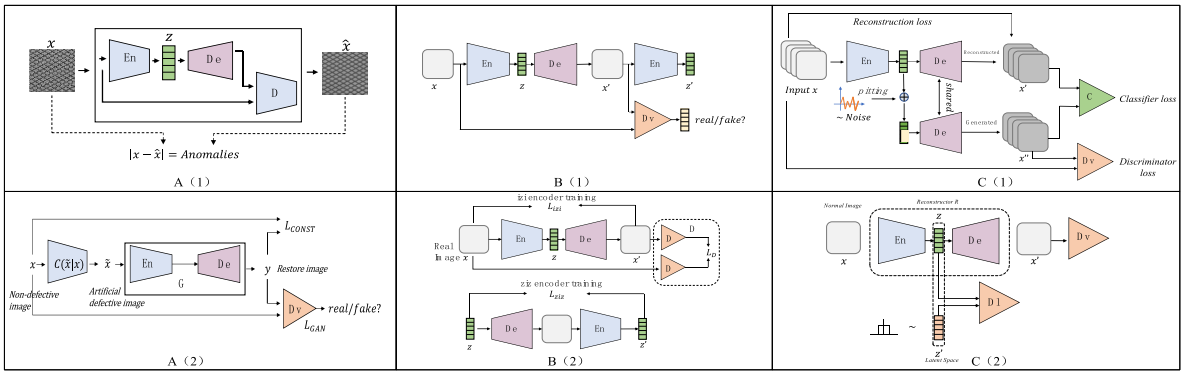

Fig. 10. Pipelines of improved GAN-based AD networks. (A1) and (A2) improving the input of the generator G. (B1) and (B2) improving the generator G. (C1) and (C2) improving the discriminator D.

TABLE III
REPRESENTATIVE WORKS OF IMPROVED GAN-BASED AL NETWORKS

Approach	Representative networks	Year	Description
improving the input	DAE-GAN [66]	2018	The input is either a defect-free image or an artificially defective image. The generator is an AE structure for image reconstruction or inpaint
improving the generator G	GANomaly [67], f-AnoGAN [22]	2018	Reconstruction of latent space variables. Generators are "encode-decode-encode" structures or multiplexed "encode-decode" structure
improving the discriminator D	DefGAN [71]	2020	Introduction of multiple discriminators to improve the generation of normal samples

or a reconstruction sample. This GAN-based AL approach was used by Balzategui *et al.* [64] to implement the quality inspection of monocrystalline solar cells and Hou *et al.* [43] to form a divide-and-assemble the framework for AL. In addition, some methods used defective synthetic samples as an input to the generator G, as shown in Fig. 10(A2). This indicates that the generator is implementing repair or inpaint in this case. Zhao *et al.* [65] established the network by repairing defect areas in the samples and then compared the input sample and the restored one to indicate the accurate anomaly regions. In particular, a denoising AE GAN is proposed by Komoto *et al.* [66], detecting defective regions by recovering a defective product image that adds an artificial defect to a defect-free product image.

2) Improving the generator G is another common approach that aims to employ constraints on reconstructing latent space features. Akcay *et al.* [67] proposed the GANomaly network, which is an additional encoder after the AE, forming an "encode–decode–encode" structure, as shown in Fig. 10(B1). The difference between the second encoder's output and the first encoder's output is employed to assess whether the input is anomalous. This similar structure is also followed by cigarette packet anomaly location [68], industrial surface detection [69], and textured surface detection [70]. Besides, Schlegl *et al.* [22] proposed another scheme, namely, f-AnoGAN, which fixes the trained decoder in the generator G and reuses it as a generator for the latent space reconstruction network. It is worth noting that the strategy of constructing the reconstruction training of feature vectors in latent space is embraced, as depicted in Fig. 10(B2).

3) Improving the discriminator D is generally made by employing multiple discriminators to enhance the discriminative ability of the GAN network. As shown in Fig. 10(C1), Zhang *et al.* [71] proposed that DefGan designs an additional branch of the reconstructed image through a latent space pitting operation and a weight sharing, which form a new discriminative loss together with the original input image. Li *et al.* [72] designed an additional latent-space discriminator by constructing a new latent-space feature through random sampling, which was fed into the designed discriminator with the latent-space feature from the original generator for discrimination, as shown in Fig. 10(C2). A summary of the past representational work, covering the structure, year, and description of the imported GAN-based AL model, is presented in Table III.

c) CycleGAN: Due to the developed GAN technology, the leverage of multi-GAN to establish mappings between different feature domains has become easier to implement. The CycleGAN framework consists of four CNNs, namely, two generators and two discriminators. While the generators try to learn the mapping between the respective domains, the discriminators try to discern between real and synthesized images within one image domain. Generally, the CycleGAN-based approach has two different domains. Yu *et al.* [73] proposed an adversarial image-to-frequency transform (AIFT) network applied in unsupervised AL of road cracks. Another transformation between the image domain of a defect-free and the image domain of a synthetic defect is a classical approach, which is also the primary way of generating defect samples in defect detection [74]. Some works, e.g., [75], [76], applied

TABLE IV
REPRESENTATIVE WORKS OF NORMALIZING FLOWS FOR AL

Method	Venue	Year	Description
DifferNet [78]	WACV, 2021	Aug. 2020	Image transform for multi-scale evaluation, vanilla normalizing flow block, AL by back-propagating the negative log-likelihood loss to the input image
CFLOW-AD [79]	WACV, 2022	Jul. 2021	Multi-scale pyramid pooling, conditional vectors concatenating with the decoder coupling layers in NF block
CS-Flow [80]	WACV, 2022	Oct. 2021	Multi-scale feature maps jointly, a fully convolutional NF with cross-connections between scales
Fastflow [28]	Arxiv, 2021	Nov. 2021	Vision transformer (ViT) as the feature extractor, 2D Flow Model

CycleGAN to rail defect detection and fiber anomalies inspection. However, the CycleGAN-based methods were validated on specific datasets and lack severe results on commonly used publicly available datasets (e.g., MVTec AD [115]). Therefore, their effectiveness needs further validation. Also, it may be noted that the lack of data samples poses a challenge in training two GAN networks well at the same time in industrial scenarios.

3) Normalizing Flow (NF): Different from previously introduced generated models that cannot estimate accurate data likelihoods, NFs [77] are neural networks that learn transformations between data distributions and well-defined densities [78]. The forward pass projects data into a latent space to calculate exact likelihoods for the data given the predefined latent distribution. Conversely, data sampled from the predefined distribution can be mapped back into the original space to generate data. For the AL task, the anomaly region is obtained by measuring the distance between the feature of the test image and the estimated distribution of defective-free images. Instead of dealing with the image directly in the VAE or GAN-based methods, NF-based methods perform AL on features. Most currently available NF-based methods first leverage pretrained networks to extract normal image features and then employ NF models to estimate the corresponding distributions accurately. The first one is DifferNet, proposed by Rudolph *et al.* [78]. This model utilizes a normalizing-flow-based density estimation of image features at multiple scales. In particular, the AL result is generated by back-propagating the negative log-likelihood loss to the input image, similar to grad-CAM. However, this framework focuses on image-level anomaly classification, which is not optimized for the localization of defects on images. The anomalous localization areas in MVTec AD do not accurately fit the ground-truth range. In 2021, three NF-based AL methods were improved in three different ways and achieved surprising results on multiple datasets. Gudovskiy *et al.* [79] designed the CFLOW-AD model based on a conditional NF framework for AL. In particular, a conditional vector using a 2-D form of conventional positional encoding (PE) is proposed, then concatenating the intermediate vectors inside decoder coupling layers with the conditional vectors. CFLOW-AD achieves new state of the art for famous MVTec AD with 98.62% AUROC and 94.60% AUPRO in localization. To boost the fine-grained representations incorporating the global and local image contexts, Rudolph *et al.* [80] proposed a fully convolutional cross-scale NF (CS-Flow) that jointly processes multiple feature maps of different scales. The convolutions in the CS-Flow block were performed at two levels, with cross-connections between scales at the second level. However, many nonanomalous backgrounds still appear in the final localization results. Recently, Yu *et al.* [28] proposed a new AL network called Fastflow, which is similar in detection principle to the previous works, except that it designs a 2-D flow based on "3 × 3" and "1 × 1" convolutions. It first utilized the visual transformer as a feature extractor for normal samples, and the features are then fed into a poststage flow model for estimating the probability distribution. This model achieves excellent results in the MVTec AD with 98.5% pixel-AUROC. In summary, a comparison of four typical flow-based AL methods is shown in Table IV.

Fig. 11. Illustration of the poor image reconstructions.

Brief Summary: Table V provides a concise overview of these three types of generative model-based approaches, as well as a brief discussion of their pros and cons. The generation effect of VAE or GAN over normal areas in the image is poor, which easily leads to false detection. At present, the best localization result is achieved by NF, which combines the deep feature embedding-based methods discussed in Section III-C.

C. Deep Feature Embedding-Based Approach

Although image reconstruction or generative models succeed in several industrial scenes, several works observed that this method typically produces incorrect reconstruction results due to the lack of feature-level discriminatory information. As depicted in Fig. 11, the reconstruction part ignores the details of the image (marked with a green rectangular box). The defective-free area at the hazelnut base is not well reconstructed, which leads to overdetection problems.

To overcome the limitation of image reconstruction or generative models, another line of research proposes to employ deep feature embedding-based methods, which are generally divided into two parts: feature extraction and anomaly estimation. Hence, the final pixel-level anomaly map is generated by

TABLE V
STRENGTHS AND WEAKNESSES OF DIFFERENT GENERATIVE MODEL-BASED AL APPROACHES

Taxonomy	Methods	Strengths	Weaknesses
VAE	Attention-based	Anomaly map is generated by derivation, not reconstruction	Localization result of the anomalous region is coarse.
	Gradient-based	Changing in loss values during training can be used as a feature to identify anomalous data	
GAN	Improving the input	Employs GAN and its modifications for enhancing the ability of image reconstruction or generation	Extensive training costs. Generator might become unstable. The generation effect of normal areas in the image is poor, which leads to false detection.
	Improving the generator G		
	Improving the discriminator D		
NF	CFLOW-AD [79]	Can estimate accurate data likelihoods for normal samples	The model needs fine design, and the design criteria are different from vanilla CNN.
	CS-Flow [80]		
	Fastflow [28]		

comparing deep embedding features from target and normal images. Feature extraction part usually chooses pretrained on large-scale databases, such as ImageNet or SSL. In particular, the NF mentioned above can also be regarded as the deep feature embedding-based method combined with the generative model. According to the paradigm of anomaly estimation, we further group these methods into "knowledge distillation-based" and "deep feature modeling-based."

1) Knowledge Distillation-Based Methods: In order to better embed the deep feature information, a student–teacher framework is leveraged here. The teacher model plays as a pretrained feature extractor, and the student model is used to estimate a scoring function for AL. Bergmann *et al.* [21] proposed the uninformed students, which is the first to employ the knowledge distillation model for anomaly location. It is characterized by employing one teacher and multiple students. Student networks are then trained to regress the output of a descriptive teacher network that was pretrained on a large dataset of patches from natural images. In particular, anomalies are localized when outputs of the student networks differ from those of the teacher network and differences in the output for different student networks. However, this model only employs the output of the last layer of the network as a feature for knowledge distillation, and the multipatch approach is adopted to localize the anomaly better, which puts a burden on computing time. To deal with the above limitation, Salehi *et al.* [26] presented a multiresolution knowledge distillation approach, where considering features from multiple intermediate layers in the distillation process leads to better use of the expert's knowledge and more significant differences compared to solely utilizing the output of the last layer. In particular, its AL map is generated by back-propagating the loss to the input, which also leads to limitations in its localization effectiveness. Wang *et al.* [81] have further extended the technique of the multiscale AL method by introducing the student–teacher feature pyramid matching (STPM) model. Their AL map is generated by directly calculating the differences between the multifeature layers of the teacher network and the student network. This model enables accurate localization results and avoids the input image's path-size setting. On the MVTec AD, it achieved 98.5% AUROC and 92.1% PRO score. Besides, some works also extend the knowledge distillation framework into AE or VAE for better reconstruction results. Chung *et al.* [82] evaluated an outlier-exposed style distillation network (OE-SDN) that mimics the mild distortions caused by an AE, termed style translation. This approach utilizes the difference between the outputs of the OE-SDN and AE as an alternative anomaly score. Dehaene and Eline [83] proposed a feature-augmented VAE (FAVAE) architecture consisting of a feature extraction module with VAE architecture, where the output of the extraction module is correlated with the multilayer output of the decoder in the VAE. It can be regarded as the knowledge distillation operation.

2) Deep Feature Modeling-Based Methods: In this pipeline, a feature space is first needed to build for the input image and then to realize the measurement or comparisons of the features by feature modeling. These tricks can be clustering, or some probability distribution fitting, or some learning models. Compared with the knowledge distillation method, it often employs one end-to-end network with no distinction between teacher and student networks. Cohen and Hoshen [25] presented an alignment-based method for detecting and segmenting anomalies inside images. It constructs a pyramid of features using a pretrained Wide-ResNet50 model and employs these feature maps to find the K nearest anomaly-free images. Defard *et al.* [23] designed a patch distribution modeling (PaDiM) framework that first generated features using a pretrained CNN and then modeled normality by applying a multivariate Gaussian distribution to each location. In the test stage, the final anomaly map is generated by measuring the Mahalanobis distance of the features at each location to a "standard feature template." Since this method was the best reported result until the advent of NF, some following works [85], [87], [88], [89] adopted this design. As the method models the fixed positions of feature maps, it is only adapted to aligned datasets. In [84], Shi *et al.* proposed AE-based feature reconstruction to replace the previous Gaussian distribution modeling strategy. However, the results of AL did not outperform PaDiM. It does not perform well on the typical datasets Tile and Wood of MVTec AD. Mishra *et al.* [85] presented VT-ADL, which combines the traditional reconstruction-based methods with the benefits of vision transformer (ViT). The input image is encoded using a ViT, and its resulting features are then fed into a decoder to reconstruct the original image. Moreover, a Gaussian mixture density network models the distribution of the transformer-encoded features to estimate the distribution of the normal data in this latent space. Its structure is complex but demonstrates

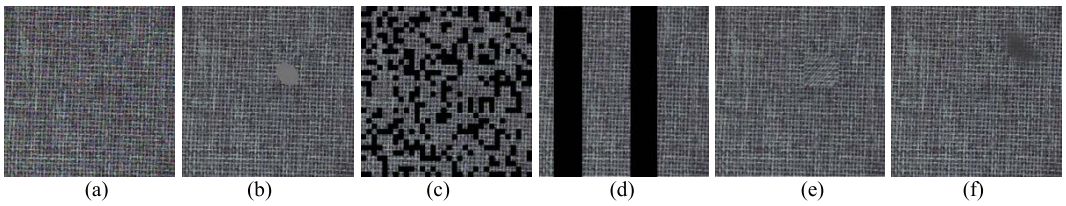

Fig. 12. Different synthetic defect images. (a) Noise. (b) Random erasure. (c) Random region mask. (d) Multiscale striped masks. (e) Cut and paste. (f) Composite synthesis.

that the visualization of the anomaly map on MVTec AD is poor. In [86], multilayer feature sparse coding (MLF-SC) is employed for AD. The AL results still rely on the image-pixel level reconstruction effect. Kim *et al.* [87] followed the idea of the PaDiM and devised a random feature selection method extending to semi-orthogonal embeddings (SOEs) to avoid the computational complexity of the multidimensional covariance tensor. It achieved good results for MVTec AD, KolektorSDD [117], and KolektorSDD2 [118]. Roth *et al.* [30] followed the idea of the Semantic Pyramid Anomaly Detection (SPADE) and proposed PatchCore for AL. It employed the nominal patch-level feature representations extracted from ImageNet pretrained networks and minimal runtime through coreset subsampling to realize a low computational cost. On MVTec AD, this method achieved more than 99% of image AD AUROC but attained inaccurate localization results. Li *et al.* [88] also followed the idea of PaDiM and used a self-organizing map (SOM) rather than a multidimensional Gaussian. This model has a slightly better AUROC at pixel level on MVTec AD than the original PaDiM. Rippel *et al.* [89] introduced fine-tuning the learned representation in the feature exacted part, thus improving the original PaDiM for AL. Yan *et al.* [90] proposed a multilevel image reconstruction and feature comparison approach to AL with an adaptive attention-to-level transformation (ALT) strategy. ALT simultaneously adjusts the weights of reconstruction levels and feature measurement scales to utilize consistent levels of features for reconstruction and AD. Tailanian *et al.* [91] designed a contrario framework to detect anomalies in images, which computed the number of false feature maps before generating the final anomaly map. However, this method only achieved AUROC of 0.77 and 0.86 on tile and wood of MVTec's texture dataset. Recently, Zheng *et al.* [92] revisited the issue of unaligned data in PaDiM. They proposed the "focus your distribution" (FYD) model, which employed a coarse to fine process. Before extracting features, an image-level coarse alignment module was designed, which allowed the input image to be forcibly aligned. Pixelwise noncontrastive learning was then utilized in the exact alignment stage, which achieved the fine alignment of dense features. This method achieved 98.2% AUROC on the MVTec AD. However, it can be seen from the AL heatmaps that there are some interference and coarse defect regions in the final result. In summary, a comparison between key elements of different state-of-the-art works is presented in Table VI. This table shows that most methods chose the Wide-ResNet50 as a pretrained model and generated the final anomaly map by the Mahalanobis distance. Besides benefiting from the pretrained feature, the deep feature modeling-based model may have more significant potential.

Brief Summary: Table VII briefly summarizes the merits and disadvantages of several deep feature embedding-based methods to AL. Table VI also includes the benefits and drawbacks of the representative's deep feature modeling. These models are attempting to address two following issues. The first one is to generate fine-grained and noise-resistant localization results. The second one is to extend the model to tackle multiscale anomalies and nonaligned datasets. We believe that it will garner increased interest from both academics and industry.

D. Self-Supervised Learning-Based Approach

SSL is the process of learning visual features from unlabeled images and then applying them to the relevant visual task. There are two SSL-based AL approaches proxy tasks and contrast learning. Proxy tasks relative focus on the development of the pretext task. On the other hand, contrast learning is primarily concerned with network design.

1) Proxy Tasks: The pretext task often takes many different forms, but it all boils down to predicting or recovering hidden regions or properties in an input image. Recent SSL-based AL methods have relied on three primary proxy tasks: image inpainting, relative position prediction, and attribute restoration.

1) Image inpainting is the most common proxy task. Self-supervision based on image inpainting is the same as previous methods based on image reconstruction or generation, only referred to differently. By repairing the defective synthetic images, the network model is given the ability to reconstruct normal sample regions and repair abnormal regions. This kind of method can repair similar abnormal regions in the test stage. A common way of synthesizing defective images is shown in Fig. 12. The earliest defect images are generated by adding random noise; for instance, Nakazawa and Kulkarni [93] used synthesized noisy images for AD in wafer images, and the multiscale AE method proposed by Mei *et al.* [37] also adopted the similar synthesized noisy images. The network model used in this approach is also known as a "denoising encoder." Besides, some data augmentation methods have been applied to generate defective training samples to improve the network's repair capability. Tayeh *et al.* [94] and Li *et al.* [95], for example, randomly erase regions in normal samples with arbitrary shapes and then fill them with a fixed color, as shown in Fig. 12(b). However, this design does not consider the structural information present in the image that facilitates subsequent network restoration.

TABLE VI
COMPARISON OF PAST WORKS WITH SOME KEY ELEMENTS

Methods	Year	Pre-trained	Normal images usage	Anomaly map	Multi-Scale	Pros	Cons
SPADE [25]	2020.5	WideResNet50	KNN	Euclidean distance	DNN scales	Simple structure, no training required	The complexity of the KNN algorithm operations is linearly related to the training samples. The more training images there are, the greater the storage requirement
PaDiM [23]	2020.11	WideResNet50	Multivariate Gaussian distribution modeling	Mahalanobis distance by fixed positions	DNN scales	Distributed estimation improves anomaly location performance	It uses separate estimation distributions at fixed locations, with significant performance degradation when the dataset is unaligned
DFR [84]	2020.12	VGG19	AE-based feature reconstruction	Reconstruction error	DNN scales	Simple learning network	Incomplete localization area. Does not perform well in many data sets, e.g., tile, cable, and transistor
VT-ADL [85]	2021.4	ViT	Gaussian mixture density network	Reconstruction error	ViT	An early introduction of ViT feature coding	The structure is complex, and the overall performance is not good
MLF-SC [86]	2021.4	VGG16	Sparse coding	Top-5 largest reconstruction errors	DNN scales	Combining sparse coding methods	Poor reconstruction result
Semi-orthogonal embedding [87]	2021.5	WideResNet50	Random feature selection, into semi-orthogonal embedding	Mahalanobis distance	DNN scales	Retaining the better performance by avoiding redundant sampling	The improvement over PaDiM is not significant and does not outperform PaDiM on tile and wood
PatchCore [30]	2021.6	WideResNet50	KNN with the memory bank	Patch distance	Patch	Faster inference and reduced feature storage capacity compared to SPADE	An improved version of SPADE, but the visualization of the anomaly maps is mediocre
SOMAD [88]	2021.7	WideResNet50	Memorizing normality via SOM	Mahalanobis distance	Patch	Self-organizing approach enhances the expression of features	Most of the anomalous regions are not refined
Gaussian fine-tune [89]	2021.8	EfficientNet -B4	Gaussian distribution	Mahalanobis distance	DNN scales	Fine-tuning of pre-trained feature representations	Visualization results are not available
MLIR [90]	2021.9	VGG19	Image reconstruction and feature comparison	Reconstruction error	DNN scales	Introduced weighting adjustment for different feature layers	The visualization of the AL results is similar to DFR
Contrario method [91]	2021.10	Resnet	Number of False Alarms computation	Mahalanobis distance	DNN scales	Applying statistical analysis to feature maps	The structure is complex, and the overall performance is not fine-grained
Focus Your Distribution [92]	2021.10	WideResNet50	Pixel-wise non-contrastive Learning	Mahalanobis distance	DNN scales	Consideration of image and feature alignment	Some of the training images are inherently unalignable, and their AD regions are not fine-grained and with a lot of interference

TABLE VII
STRENGTHS AND WEAKNESSES OF DIFFERENT DEEP FEATURE EMBEDDING-BASED AL APPROACHES

Taxonomy	Strengths	Weaknesses
Knowledge distillation	AL problem is transformed into a direct feature comparison between different networks	Easy to be disturbed by the choosing layer for knowledge distillation.
Deep feature modeling	Rich semantic information is introduced through the pre-trained model	Memory requirements are relatively high, and the feature modeling needs careful design, which greatly impacts localization results.

Therefore, Zavrtanik *et al.* [53] designed a mesh-like random mask, as shown in Fig. 12(c). The number and scale of mask regions are parameterized. On the MVTec AD, the method achieved 94.2% of the pixel AUROC. Yan *et al.* [96] presented a multiscale strip mask for modeling large span defects of different scale sizes, as shown in Fig. 12(d). Some recent works have attempted to generate realistic defective images rather than just using meaningless black and white block images. Li *et al.* [97] were the first to crop the region on the original defect-free image and then paste it onto the image at a random angle to form a new anomaly image, as shown in Fig. 12(e). Song *et al.* [98] also follow this idea. In particular, some methods leverage a more sophisticated approach to background fusion, where defects are simulated by selecting different background images, varying size, brightness, and shape, and then adding image fusion to produce a more realistic defective image, as shown in Fig. 12(f). For example, Schlüter *et al.* [99] used Poisson fusion, Zavrtanik *et al.* [29] selected various texture images as defective backgrounds, and Haselmann and Gruber [100] borrowed from sample synthesis

TABLE VIII
STRENGTHS AND WEAKNESSES OF DIFFERENT SSL-BASED AL APPROACHES

Taxonomy	Methods	Strengths	Weaknesses
Proxy tasks	image inpainting	Synthetic or simulated abnormal data	A gap between the simulated anomaly and the real anomaly
	relative position prediction	Consider the spatial information of the neighborhood patch	Correlation between neighborhood patches does not exist in many images, such as complex objects
	attribute restoration	Using image attributes, such as color and orientation	Effect of attribute and direction on anomaly location is limited
Contrast learning	CPC [105]	Using similarity to distinguish between normal and abnormal	Localization result is easily affected by interference, such as variations in imaging conditions
	SimSiam [106]		
	SimCLR [108]		

methods in data augmentation. Theoretically, the closer to the effect of a real synthetic defect, the more generalizable the image reconstruction and restoration capability should be. However, in real scenes, the type and shape of defects are often unpredictable, so it is difficult to decide which of the synthesis methods is optimal. As shown in Table X, there is no relationship between the more realistic defect synthesis method and good positioning results. In general, these methods often need to be combined with designing a suitable restoration network to achieve better results.

2) *Relative location prediction:* Different from previously introduced models, which only consider the mapping between input and output, there is another way to assess the spatial information of the neighborhood patch. The most representative method is Patch Support Vector Data Description (SVDD) [24], which introduces a self-supervised approach to feature extraction. It first divides the image into 3×3 patch regions, and the eight blocks around the central image block are sorted in order. Then, the encoder of this model is trained to extract informative features, so that the following classifier can correctly predict the relative positions of the patches. However, due to the patch region setting, the AL results for this type of method are often very rough and not fine-grained. Pirnay and Chai [27] designed an InTra network based on image restoration. Specifically, it aims at a patch region centered on $w \times w$ that can be restored by the image information of its surrounding patches, so it also leverages neighborhood information. Ristea *et al.* [101] designed a self-supervised predictive convolutional attentive block (SSPCAB) for AL. For each location where the dilated convolutional filter is applied, the block learns to reconstruct the masked area using contextual information. Note that this approach is essential to employ the regional features of the dilated convolution to model a more extensive range of neighborhoods.

3) Attribute restoration is characterized by using hidden attributes in the image rather than masked areas. These attributes generally include color and orientation. Fei *et al.* [102] proposed an attribute restoration network that turns the traditional reconstruction task into a restoration task. It first changes specific attributes of the input (e.g., removes color, changes orientation, and so on) before feeding the image into AE for reconstruction. Ulutas *et al.* [103] presented a split-brain CAE approach to detect and localize defects. Two disjointed CAE networks are employed to predict the subchannel of the image from another subchannel. Each encoder implements a conversion between different color channels. This design utilizes the color property and boosts the localization accuracy of the anomaly image.

2) Contrast Learning: As described before, the proxy task focuses on generating images similar to the training data at the pixel level. Another improvement is learning common features between similar instances and distinguishing differences between nonsimilar instances. De Haan and Löwe [104] directly applied contrastive predictive coding (CPC) [105] to detect and segment anomalies in images. It splits the image into patches, interpreting each line of patches as a separate time step. In the test stage, the test image block is compared with a randomly selected image block in a defect-free image to calculate the contrast loss function, namely, InfoNCE. The current image block is judged as an abnormal region when a certain threshold is exceeded. As a result, this method affects the detection efficiency due to the patch-based operation, and its localization accuracy is not high. In [92], the fine alignment part in the proposed network for AL is designed based on SimSiam [106]. It inputs the results of two random transformations of the same feature, extracts the features employing the same encoder f, and transforms them to a higher dimensional space. The predictor g is employed, which transforms the result of one of the branches and matches it with the result of the other branch. This approach takes full advantage of the Siamese network's natural modeling invariance. Gui [45] followed the same idea of Siamese architecture for AL, except that the original predictor was replaced with a self-supervised module. Yoa *et al.* [107] presented an AL method based on simple framework for contrastive learning of visual representations (SimCLRs) [108]. By generating a pair of negative images in the training dataset, the design model contrasts a normal sample to a locally augmented sample. This model achieved 93.4% pixel-AUROC on the MVTec AD. SSL frameworks are still a hot topic of research, and we believe that these novel models can show and validate AL, which will constitute a relevant future direction.

Brief Summary: Table VIII summarizes the two types of SSL-based approaches and their benefits and drawbacks. SSL frameworks are still a hot issue of research in general. We believe that these novel models show and validate the potential of the AL and constitute a relevant future direction.

E. One-Class Classification-Based Approach

The one-class classification approaches are usually employed for image-level AD, typical including one-class support vector machine (OCSVM) [109] and deep SVDD [110]. Deep SVDD trains a network and then maps the training data to a small hypersphere in the feature space. The data outside the hypersphere are called anomalies. For AL, a one-class classification-based approach locates anomalous regions by dividing the image into patches and classifying the patches into abnormal or normal categories, which enables coarse results. This form was adopted by Liu *et al.* [111] and Wang *et al.* [112] to localize anomalous regions on the steel surfaces and wind turbine blades, respectively. Furthermore, several improved versions of deep SVDD have been proposed for AL. Compared to deep SVDD, Patch SVDD [24] inspects every patch to localize a defect, and SSL is employed, allowing the features to form multimodal clusters, thereby enhancing AD capability. In addition, the deep SVDD method was embedded in the pretrained feature comparison by Hu *et al.* [113]. It estimates the pixelwise anomalies efficiently based on deep SVDD. Liznerski *et al.* [114] presented a fully convolutional data description (FCDD), a modification of deep SVDD, so that the transformed samples are themselves an image corresponding to a downsampled anomaly heatmap. While this approach produces a full resolution anomaly heat map, the extent of the anomaly region is not accurate due to the upsampling operation of a fixed Gaussian kernel.

Brief Summary: Major AD approaches can be employed for pixel-level AL, since we can segment the complete image into several patches and then perform AD on image patches. The AD algorithm concentrates on the entire image's semantic information; therefore, the semantic information of subtle abnormal regions may be ignored. Here, we observe that combining AD with some self-supervised strategies or pretrained deep feature embedding methods is a promising way to increase localization performance.

IV. Experiments

A. Datasets Used by Recent Works

Five datasets are available for unsupervised learning-based AL datasets, which differ significantly in terms of image quantity, quality, resolution, and texture information.

NanoTWICE [16] is the first dataset proposed to apply the AD problem. It contains 45 nanofibrous material images with 1024×3696 pixels captured from a scanning electron microscope. The background of the image is a noncyclical continuous texture, and the size of the defect varies.

MVTec AD [115] is currently the most common dataset for industrial AL, which contains 15 categories; each category has about 240 normal images for training and 100 defective images for testing. The original image resolution is between 700×700 and 1024×1024 pixels. Compared with the existing datasets that focus on texture defects, this dataset has ten objects and five texture types. The five categories cover different types of regular (carpet and grid) or random (leather, tile, and wood) textures, while the remaining ten categories represent various types of objects. Some of these objects are rigid and have a fixed appearance (bottles and metal nuts), while others are deformable (cables) or include natural changes (hazelnuts). The test images of abnormal samples contain various defects, such as scratches, dents, structural differences, and so on. There are a total of 73 different types of defects, with an average of about five for each category. More details of this dataset can be found in [115]. However, this dataset is well imaged and uniformly illuminated, while the image positions are fixed in some data types, making it a more idealized.

BeanTech anomaly detection (BTAD) dataset has been recently released by Mishra *et al.* [85]. It contains 2830 images with three different classes. The resolutions of these three classes are 1600×1600, 600×600, and 800×600 pixels, respectively. Each class is composed of defect-free training and testing images, such as the MVTec AD dataset, except that the defect types are not illustrated.

Fabric dataset [116] is from the automation laboratory sample database of Hong Kong University constructed by Tsang *et al.*, which contains 256×256 fabric images belonging to three patterns: dot, star, and box-patterned fabrics. Each pattern has 25 defect-free and 25 defective samples. There are five types of defects that appear in the defective samples, which include the broken end, hole, netting multiple, thick bar, and thin bar. All the defective fabric images have the corresponding ground truth. This dataset is a classical textured dataset, which is often employed in fabric defect detection works.

Textured dataset is also created by MVTec company and first presented in the AE-SSIM [18]. This dataset contains two woven fabric textures. All images are of size 512×512 pixels that were acquired as single-channel gray-scale images.

Besides these, some datasets widely used for other supervised industrial vision tasks are also employed for AL, such as KolektorSDD [117], KolektorSDD2 [118], railway surface discrete defects (RSDDs) [119], and magnetic tile (MT) defect [120] datasets. Typically, the defect-free samples of the training set in these datasets are used for AL model training. Then, the remaining samples, mainly defective, are utilized as test sets.

KolektorSDD [117] and KolektorSDD2 [118] are the datasets of metal surfaces collected in real industrial scenarios. KolektorSDD is relatively simple, with only one thin scratch defect. KolektorSDD2 is a real-world complex and well-annotated modern surface inspection dataset. It is constructed from color images of defective production projects, captured using a visual inspection system, and annotated by the company. The defects are annotated with fine-grained segmentation masks that vary in shape, size, and color, ranging from minor scratches to spots on large surface defects. The RSDD [119] and MT Defect [120] datasets are also frequently employed to evaluate AL. The RSDD dataset contains two types of datasets collected on real railroad tracks, and its texture and illumination vary significantly. The MT Defect dataset includes five types of defects: blowhole, break, uneven, fray, and crack, all of which have different resolutions. These defect images contain a series of industrial noises, such as the changes in light intensity, the defect's scale, and the texture's

TABLE IX
COMMON DATASETS FOR INDUSTRIAL ANOMALY LOCALIZATION

Name	URL*	Description	Flaw
NanoTWICE [16]	http://www.mi.imati.cnr.it/ettore/NanoTWICE	45 nanofibrous material images of size 1024 × 3696, non-cyclical continuous texture, only textural anomalies	Single continuous texture, too few samples
MVTec AD [115]	https://www.mvtec.com/company/research/datasets/mvtec-ad	15 categories, with about 240 normal images for training and 100 defective images for testing in each category, Image size ranges from 700×700 to 1024×1024, textural anomalies and functional anomalies	Too consistent image illumination, serious alignment in some data sets, and too few functional anomalies
BTAD [85]	http://avires.dimi.uniud.it/papers/btad/btad.zip	2830 images with 3 different classes, only textural anomalies	Serious alignment, too few defect types, too consistent image illumination
Fabric dataset [116]	https://ytngan.wordress.com/codes	3 categories, each having 25 defect-free and 25 defective samples, only textural anomalies	Uniform texture background, too consistent image illumination,
Textured dataset [18]	https://www.mvtec.com/company/research/publications	2 woven fabric textures with 512 × 512 pixels, only textural anomalies	Uniform texture, consistent image illumination
KolektorSDD [117]	https://www.vicos.si/resources/kolektorsdd	347 images without defects and 52 images with defects, only textural anomalies	Only with one defect type-thin scratch
KolektorSDD2 [118]	https://www.vicos.si/resources/kolektorsdd2	Over 3000 images containing several types of defects, rich defect scales, only textural anomalies	Uniform texture
RSDDs [119]	http://icn.bjtu.edu.cn/Visint/resources/RSDDs.aspx or, https://pan.baidu.com/s/1kM5Lh9-s2yImuQE2ks0dqg (password: lvth)	Images with sizes 160×160 and 55×55, only textural anomalies	Too few image samples and constant defect size
MT Defect [120]	https://github.com/abin24/Magnetic-tile-defect-datasets.	5 types of metal defects, rich imaging scenes, only textural anomalies	A very few samples in each category, noise in labels

* last accessed: 6th May, 2022

complexity, but these do not contain many types and variations of defects. In Table IX, we list multiple image datasets commonly used by the AL community and specifically indicate their download links, descriptions, and flaws.

B. Evaluation Criteria

AUROC, per region overlap (PRO) score, and intersection over union (IoU) are the three primary evaluation metrics used in AL, as discussed below.

1) Receiver Operating Characteristic Curve (AUROC): The most frequent indication of AL is the AUROC of each pixel. The high AUROC value indicates that the model is less influenced by varied threshold settings while identifying anomalies. Normal pixels are identified as negative, whereas anomalous pixels are identified as positive. The true positive rate (TPR) is the percentage of pixels properly categorized as anomalous across the evaluated category, whereas the false positive rate (FPR) is the percentage of pixels incorrectly classified as abnormal, which is denoted as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{1}$$

where TP, FP, TN, and FN denote the true positive, false positive, true negative, and false negative, respectively.

The AUROC value may be determined by scanning across the range of thresholds and obtaining sorted serial values of TPR and FPR. However, in surface AD settings where only a tiny proportion of pixels are anomalous, the AUROC does not accurately reflect localization accuracy. The reason is that the FPR is dominated by the very high number of nonanomalous pixels and is thus kept low despite false-positive detection. Therefore, despite achieving about 97% pixel AUROC, some state-of-the-art methods cannot produce fine-grained AL results. They often have more interference and introduce too much background area, as seen in the visualized anomaly maps. Despite this, AUROC is currently the dominant evaluation indicator used.

2) Per Region Overlap (PRO) Score: Since the AUROC favors large anomalies, the PRO score is also employed for AL. For computing the PRO metric, anomaly scores are first thresholded to make a binary decision for each pixel whether an anomaly is present or not. For each connected component within the ground truth, the relative overlap with the thresholded anomaly region is computed. Numerous approaches [21], [23], [50], [84], [88], [116] also use the PRO score to evaluate the performance of the model.

3) Intersection Over Union (IoU): AL can be treated as a segmentation task similar to that in supervised learning. The IoU, as the primary metric for segmentation tasks, can equally justly be used to evaluate the performance of AL. Currently, only very few works use this evaluation approach, such as [49], [59], and [98].

C. Performance Comparison

1) Performance on MVTec AD Datasets: Tables X and XI summarize the performance of the contemporary AL methods (published mainly from the year 2017 to 2021) on the MVTec AD dataset. We observed that most of these approaches achieved baseline performance with assistance from AE. A few attempts were dedicated to design more powerful modules, such as image inpainting and GAN. The pixel AUROC on the MVTec AD dataset has been up to 94.2% achieved by reconstruction-by-inpainting-based anomaly detection (RIAD) [53]. Nevertheless, experimental results

TABLE X

LOCALIZATION RESULTS (PIXEL AUROC %) OF THE STATE-OF-THE-ART AL METHODS ON MVTEC AD

Category	Method	Mean	carpet	grid	leather	tile	wood	bottle	cable	capsule	hazelnut	metal nut	pill	screw	toothbrush	transistor	zipper
AE-based	AESc [36]	86.0	91.0	95.0	87.0	79.0	84.0	88.0	84.0	93.0	89.0	62.0	85.0	95.0	93.0	78.0	90.0
	TrustMAE [41]	93.9	98.5	97.5	98.1	82.5	92.6	93.4	92.9	87.4	98.5	91.8	89.9	97.6	98.1	92.7	97.8
	PixelSail [47]	94.0	94.0	99.0	87.0	99.0	88.0	95.0	95.0	93.0	95.0	91.0	95.0	96.0	97.0	91.0	98.0
	UTAD [49]	90.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	MPAD [50]	98.1	98.4	98.5	99.1	94.4	97.5	98.6	98.2	97.9	97.8	99.1	98.8	98.5	99.0	97.7	98.6
	MIGD [51]	91.0															
	AE-SSIM [18]	86.2	87.0	94.0	78.0	59.0	73.0	93.0	82.0	94.0	97.0	89.0	91.0	96.0	92.0	80.0	88.0
	AE-L_2 [18]	82.0	59.0	90.0	75.0	51.0	73.0	86.0	86.0	88.0	95.0	86.0	85.0	96.0	93.0	86.0	77.0
	RIAD [53]	94.2	96.3	98.8	99.4	89.1	85.8	98.4	84.2	92.8	96.1	92.5	95.7	98.8	98.9	87.7	97.8
Generative model-based	Vanilla VAE [55]	82.3	62.0	85.6	83.5	52.0	69.9	89.4	81.6	90.7	95.1	86.1	87.9	92.8	95.3	85.1	77.5
	VE-VAE [57]	86.1	78.0	73.0	95.0	80.0	77.0	87.0	90.0	74.0	98.0	94.0	83.0	97.0	94.0	93.0	78.0
	CAVAG [59]	89.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	VAE-grad [61]	89.0	74.0	96.0	93.0	65.0	84.0	92.0	91.0	92.0	98.0	91.0	93.0	95.0	98.0	92.0	87.0
	AnoGAN [17]	74.3	54.0	58.0	64.0	50.0	62.0	86.0	78.0	84.0	87.0	76.0	87.0	80.0	90.0	80.0	78.0
	CFLOW-AD [79]	**98.6**	99.3	99.0	**99.7**	98.0	96.7	99.0	97.6	99.0	99.0	98.6	99.0	98.9	98.9	98.0	99.1
	Fastflow [28]	98.5	**99.4**	98.3	99.5	96.3	97.0	97.7	98.4	**99.1**	99.1	98.5	**99.2**	99.4	98.9	97.3	98.7
Deep feature embedding-based	S-T [21]	93.9	93.5	89.9	97.8	92.5	92.1	97.8	91.9	96.8	98.2	97.2	96.5	97.4	97.9	73.7	95.6
	MKD [26]	90.7	95.6	91.7	98.0	82.7	84.8	96.3	82.4	95.9	94.6	86.4	89.6	96.0	96.1	76.5	93.9
	STPM [81]	97.0	98.8	99.0	99.3	97.4	97.2	98.8	95.5	98.3	98.5	97.6	97.8	98.3	98.9	82.5	98.5
	OE-SDN [82]	93.0	96.0	97.0	85.0	85.0	82.0	95.0	84.0	97.0	98.0	93.0	93.0	97.0	98.0	89.0	91.0
	FAVAE [83]	95.3	96.0	99.3	98.1	71.4	89.9	96.3	96.9	97.6	98.7	96.6	95.3	99.3	98.7	**98.4**	96.8
	SPADE [25]	96.5	97.5	93.7	97.6	87.4	88.5	98.4	97.2	99.0	99.1	98.1	96.5	98.9	97.9	94.1	96.5
	PaDiM [23]	97.5	99.1	97.3	99.2	94.1	94.9	98.3	96.7	98.5	98.2	97.2	95.7	98.5	98.8	97.5	98.5
	DFR [84]	95.0	97.0	98.0	98.0	87.0	93.0	97.0	92.0	99.0	99.0	93.0	97.0	99.0	99.0	80.0	96.0
	SOE [87]	98.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	PatchCore [30]	98.1	99.0	98.7	99.3	95.6	95.0	98.6	98.4	98.8	98.7	98.4	97.4	99.4	98.7	96.3	98.8
	SOM [88]	97.8	98.9	98.4	99.1	94.8	94.4	98.3	98.2	98.7	98.4	98.0	98.0	99.1	98.5	95.3	98.7
	TF [89]	96.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	ALT [90]	96.9	97.1	99.5	98.9	95.5	96.2	96.4	90.8	98.8	99.1	97.6	98.5	99.3	97.7	91.4	97.5
	FYD [92]	98.2	98.5	96.8	99.2	96.8	**99.6**	98.3	97.5	98.6	98.7	98.2	97.3	98.7	98.9	98.1	98.2
Self supervised learning-based	SMAI [95]	89.0	88.0	97.0	86.0	62.0	80.0	86.0	92.0	93.0	97.0	92.0	92.0	96.0	96.0	85.0	90.0
	Cutpaste [97]	96.0	98.3	97.5	99.5	90.5	95.5	97.6	90.0	97.4	97.3	93.1	95.7	96.7	98.1	93.0	**99.3**
	ANOSEG [98]	97.0	99.0	99.0	98.0	98.0	98.0	99.0	**99.0**	90.0	99.0	99.0	94.0	91.0	96.0	96.0	98.0
	NSA [99]	96.3	95.5	99.2	99.5	**99.3**	90.7	98.3	96.0	97.6	97.6	98.4	98.5	96.5	94.9	88.0	94.2
	DRAEM [29]	97.3	95.5	**99.7**	98.6	99.2	96.4	**99.1**	94.7	94.3	99.7	**99.5**	97.6	97.6	98.1	90.9	98.8
	InTra [27]	96.6	99.2	98.8	99.5	94.4	88.7	97.1	91.0	97.7	98.3	93.3	98.3	99.5	98.9	96.1	99.2
	SSPCAB [101]	97.2	95.0	99.5	99.5	**99.3**	96.8	98.8	96.0	93.1	**99.8**	98.9	97.5	**99.8**	98.1	87.0	99.0
	CPC-AD [104]	82.0	74.0	80.0	94.0	82.0	82.0	89.0	84.0	72.0	81.0	76.0	77.0	65.0	81.0	90.0	95.0
	DLA [107]	93.0	89.4	88.1	98.5	91.9	89.2	91.8	88.3	96.5	96.2	92.6	96.4	97.2	95.8	88.3	95.4
One-class classification-based	Patch-SVDD [24]	95.7	92.6	96.2	97.4	91.4	90.8	98.1	96.8	95.8	97.5	98.0	95.1	95.7	98.1	97.0	95.1
	SE-SVDD [113]	97.5	98.9	97.2	98.7	92.3	95.1	98.6	97.7	98.5	98.0	98.3	96.7	98.6	**99.3**	97.2	97.9
	FCDD [114]	92.0	96.0	91.0	98.0	91.0	88.0	97.0	90.0	93.0	95.0	94.0	81.0	86.0	94.0	88.0	92.0

In each column, the best result is marked **bold**

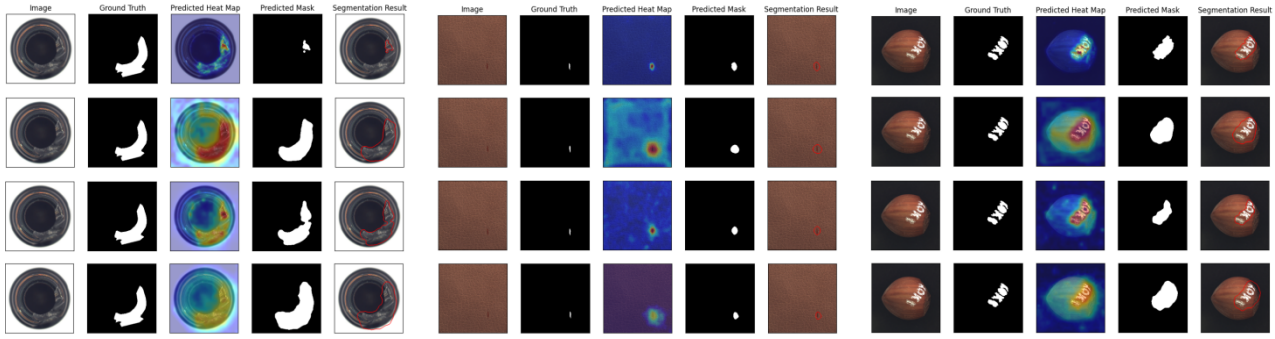

Fig. 13. Qualitative results on MVTec AD samples of STPM [81], PatchCore [30], PaDiM [23], and CFLOW-AD [79], rowwise.

demonstrated that these pure AE-based reconstructions or generative methods are hardly capable of performing sufficiently well on the MVTec AD dataset.

In contrast, deep feature embedding-based methods have quickly demonstrated their strengths in AL. The reported results in the past papers show that three typical feature compared methods, S-T [21], SPADE [25], and deep feature reconstruction (DFR) [84], and achieved overall 93.9%, 96.5%, and 95.0% pixel AUROC on the MVTec AD dataset, respectively. Starting from the generic feature modeling method [23], feature embedding-based methods improve steadily when introduced with more effective strategies, e.g., feature selection into SOE [87], attention strategy [23], [43], k-Nearest Neighbor (KNN) with the memory bank [30], self-organizing feature [88], and aligning feature [92]. As a result, most approaches yielded about a pixel AUROC of 93% and a PRO score of 91% on the MVTec AD dataset. Moreover, CFLOW-AD [79], combined with a novel generative network, outperformed other state-of-the-art models and achieved the best pixel AUROC on MVTec AD so far. On the other hand, MPAD [50], combined with pretrained features, surpassed other state-of-the-art models and achieved the best PRO score on MVTec AD so far. Here, in Fig. 13, we present the visualization of AL results of four typical feature embedding methods over MVTec

TABLE XI

LOCALIZATION RESULTS (PRO SCORE %) OF THE STATE-OF-THE-ART AL METHODS ON MVTEC AD

Category	Method	Mean	carpet	grid	leather	tile	wood	bottle	cable	capsule	hazelnut	metal nut	pill	screw	toothbrush	transistor	zipper
AE-based	AE-SSIM [18]	69.4	64.7	84.9	56.1	17.5	60.5	83.4	47.8	86.0	91.6	60.3	83.0	88.7	78.4	72.4	66.5
	PixelSail [47]	50.0	47.0	89.0	80.0	36.0	53.0	52.0	40.0	31.0	54.0	36.0	24.0	47.0	69.0	8.0	82.0
	MPAD [50]	**95.5**	92.7	97.9	**99.2**	88.8	**96.2**	95.3	**96.7**	**97.8**	97.8	88.8	96.1	98.3	94.4	95.0	97.0
Generative model-based	Vanilla VAE [55]	64.2	61.9	40.8	64.9	24.2	57.8	70.5	77.9	77.9	77.0	57.6	79.3	66.4	85.4	61.0	60.8
	CFLOW-AD [79]	94.6	**97.7**	96.1	99.4	94.3	95.8	**96.8**	93.5	93.4	96.7	91.7	95.4	95.3	95.1	91.4	96.6
Deep feature embedding based	S-T [21]	91.4	87.9	95.2	94.5	94.6	91.1	93.1	81.8	96.8	96.5	94.2	96.1	94.2	93.3	66.6	95.1
	STPM [81]	92.1	95.8	96.6	98.0	92.1	93.6	95.1	87.7	92.2	94.3	94.5	**96.5**	93.0	92.2	69.5	95.2
	SPADE [25]	91.7	94.7	86.7	97.2	75.9	87.4	95.5	90.9	93.7	95.4	94.4	94.6	96.0	93.5	**97.4**	92.6
	PaDiM [23]	92.1	96.2	94.6	97.8	86.0	91.1	94.8	88.8	93.5	92.6	85.6	92.7	94.4	93.1	84.5	95.9
	DFR [84]	91.0	93.0	93.0	97.0	79.0	91.0	93.0	81.0	97.0	97.0	90.0	96.0	96.0	93.0	79.0	90.0
	SOE [87]	94.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	PatchCore [30]	93.5	96.6	95.9	98.9	87.4	89.6	96.1	92.6	95.5	93.9	91.3	94.1	97.9	91.4	83.5	**97.1**
	SOM [88]	93.3	95.5	95.3	97.7	81.3	88.2	94.7	93.4	93.4	95.1	93.6	96.5	96.0	90.7	91.6	95.9
	TF [89]	88.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	ALT [90]	94.0	91.3	**99.0**	96.8	90.1	92.5	94.3	88.9	97.1	**98.0**	90.3	94.0	**98.9**	**96.3**	88.4	94.0
	FYD [92]	91.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Self supervised learning based	NSA [99]	91.0	85.0	96.8	98.7	**95.3**	85.3	92.9	89.9	91.4	93.6	**94.6**	96.0	90.1	90.7	75.3	89.2
One-class classification based	SE-SVDD [113]	92.3	96.1	94.3	96.2	87.5	90.7	93.9	87.9	93.3	93.7	93.5	93.2	93.3	93.1	85.5	93.7

In each column, the best result is marked **bold**

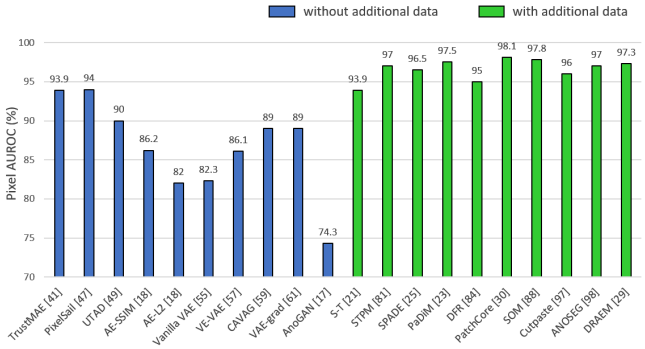

Fig. 14. Comparison among the state-of-the-art AL methods with/without additional data on the MVTec AD dataset.

AD, including STPM [81], PatchCore [30], PaDiM [23], and CFLOW-AD [79]. These results were obtained using a standard image AL library Anomalib [125], maintained by Intel corporation.

SSL-based methods can learn visual features from unlabeled images and be embedded into the above network structure as an additional module. Such methods, e.g., anomaly segmentation network (ANOSEG) [98], Natural Synthetic Anomalies (NSA) [99], and discriminatively trained reconstruction anomaly embedding model (DRAEM) [29], can achieve better results compared with original AE-based methods. Moreover, contrast learning-based methods [92], [107] demonstrate very competitive performance due to the discriminative information of the anomaly regions, compared to image reconstruction or pretrained features. One-class classification-based methods are usually time-consuming and obtain inaccurate localization results, especially the computational time of cropping local patches and extracting individual local features. However, some methods include more complicated feature comparison procedures, e.g., Patch-SVDD [24] and SE-SVDD [113] are designed in a unified pipeline to improve the localization performance.

In summary, the deep learning-based AL methods can obtain a relatively satisfactory result on the MVTec AD dataset by adopting different strategies. In particular, three datasets among the 15 datasets have not been overcome by the majority of methods; those are the tile, wood, and transistor datasets. Tile and wood are typical texture datasets that contain multiscale and multitype defects, and major methods currently do not achieve 95% AUROC. The transistor dataset has a missing defect type containing high-level semantic information. In this dataset, it treats all ranges of the missing as ground truth. Hence, major current methods also do not achieve ideal performances.

2) CNN Versus ViT Versus NF: We also analyze the performance of the MVTec AD dataset using different network modules under a deep feature embedding-based approach. Results in pixel AUROC (%) of representative methods are shown in Table XII. The two best algorithms, CFLOW-AD [79] and Fastflow [28], both employed the NF module training method. U-TRansformer based Anomaly Detection (UTRAD) [124] and InTra [27] are two approaches that leveraged ViT by utilizing transformer layers to build the reconstruction model. Other algorithms used a simple CNN module. ViT can capture a wide variety of visual areas through an attention mechanism, whereas NF directly estimates the probability of a normal sample. Table XII comprehends that combining ViT or NF can significantly improve the localization accuracy. As a matter of fact, more attention is encouraged to use NF and ViT in the future.

3) Additional Data Versus Without Additional Data: AL in industrial scenarios is a very challenging problem. It is difficult to achieve good results by relying solely on image-level data reconstruction or generation. Most existing approaches take assistance from additional data used in large pretrained models, data synthesis by self-supervised means, or designing proxy tasks. In Fig. 14, we present a bar graph for easier performance comparison of models with and without additional data usage over MVTec AD. It can be observed that major methods employing additional data outperformed the methods that did not use extra data.

4) Run-Time Analysis: The real-time characteristic of the AL in industrial images is a special feature. Hence, in Table XIII, we provide the run-time analysis of some major

TABLE XII
COMPARISON OF DIFFERENT NETWORK MODULES ON MVTEC-AD

Methods		Venue	Key points	Performance
Vanilla CNN	S-T [21]	IJCV-2021	Knowledge distillation	93.9
	MKD [26]	CVPR-2021	Knowledge distillation	90.7
	SPADE [25]	ArXiv-2020	Feature modeling	96.5
	DFR [84]	NeuroComp2021	Feature modeling	95.0
	FAVAE [83]	ArXiv-2020	Feature modeling	95.3
ViT	InTra [27]	ArXiv-2021	Image inpainting	96.6
	UTRAD [124]	Neural Net-2021	Feature modeling	96.7
NF-based	CFLOW-AD [79]	WACV-2022	Feature modeling	**98.6**
	Fastflow [28]	ArXiv-2021	Feature modeling	98.5

In the last column, the best result is marked **bold**

TABLE XIII
INFERENCE SPEED (fps) OF VARIOUS AL METHODS ON MVTEC-AD

Methods	Inference speed (fps)
AnoGAN [17]	0.02
GANomaly [67]	9.1
Skip-GANomaly [35]	7.9
DifferNet [78]	2.04
PaDiM-Resnet18 [23]	4.4
SPADE-WRN50 [25]	0.1
CFLOW-AD-WRN50 [79]	**27**
DFR [84]	20
Patch Core-WRN50 [30]	5.88
FastFlow-WRN50 [28]	21.8

In the last column, the best result is marked **bold**

AL methods. As shown in this table, the AL methods based on deep feature embedding touch 20 fps. In particular, CFLOW-AD-WRN50 [79] achieved 27 fps, demonstrating its ability for defect detection in real time.

V. CONCLUSION AND OUTLOOK

This article has highlighted recent achievements in industrial AL using deep learning. Here, we also provided some structural taxonomy for various methods based on their roles for AL, analyzed their advantages and limitations, summarized existing popular industrial AL datasets, and discussed performance for the most representative approaches. Despite significant progress, several issues remain unresolved. This section will highlight these issues and present some possible future research directions. We anticipate that this study not only improves an academic understanding of industrial AL but also encourages future research efforts.

A. Functional Anomalies

From the strengths and weaknesses mentioned in the above tables, it can be observed that the localization effects of many methods drop significantly on some specific datasets. For example, the disadvantage of DFR [84] is the poor performance on transistor datasets (refer to Tables VI and X). This is because most of the datasets shown in Table X are textural defects, such as scratches and dents, rather than functional anomalies. Functional anomalies violate underlying constraints, e.g., a permissible object in an invalid location or the absence of a required object. In industrial scenes, both types are equally important. At present, there is already a method by Bergmann *et al.* [126] to jointly detect the textural and functional anomalies. However, the research on functional defects will be an important direction in the future.

B. Releasing Rich AL Datasets

Compared to real industry scenarios, public anomaly location datasets are not yet large or rich enough. More complex datasets with changing imaging conditions, such as lighting, perspective, proportion, shadow, blur, and so on, should be available to evaluate the effect of the AL algorithm more objectively. The existing MVTec AD has single imaging, relatively good image quality, and alignment in some categories. Some existing AL methods even exploit this property for performance enhancement. Despite the promising results achieved, these methods cannot be adapted to real complex industrial scenarios. Therefore, it is necessary to have some realistic and rich industrial AL datasets.

C. Vision Transformer-Based Method

The ViT-based methods currently dominate the field of computer vision since their superior performance. Some ViT-based works [27], [124], [79] have also been proposed to solve the AL problem. ViT has particular advantages in long-distance feature modeling. Comprehensively considering multiscale anomalous regions is a direction that ViT can improve. Moreover, the best framework for AL is the generation model based on NF. Therefore, the combination of ViT and NF has also been an important direction.

D. Meaningful Model Evaluation

As stated above, there is an overlap between the high pixel-AUROC value and the fine-grained localization performance, which may cause the model validity problem. Many methods still utilize the pixel-AUROC evaluation metric, but the visualization results of the AL do not perform well. Future works are recommended to consider the problem of fine boundary when building their models or choose the IoU metric for model evaluation.

E. Accurate Anomaly Types

The types of anomalies in real industrial scenarios are diverse, and the importance of different anomaly types varies. This problem challenges the classical paradigm of AD or localization and requires the development of learning methods that can discriminate between anomaly types. There are already methods [122] to cluster anomaly types and group anomaly data into semantically consistent categories, but this is only a start.

1) Unsupervised 3-D Anomaly Localization: With the spread of 3-D sensors, an increasing number of defect detection tasks in industrial scenarios are moving from 2-D to 3-D scenarios. Correspondingly, AL in 3-D scenes will be a trend for the sake of development. Recently, MVTec company made a 3-D AD/AL dataset publicly available in late 2021 [123]. Therefore, we believe that 3-D AD/AL constitutes a relevant future direction.

REFERENCES

[1] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 50, no. 4, pp. 1486–1498, Apr. 2020.

[2] X. Tao, D. Zhang, W. Hou, W. Ma, and D. Xu, "Industrial weak scratches inspection based on multifeature fusion network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[3] X. Tao *et al.*, "Wire defect recognition of spring-wire socket using multitask convolutional neural networks," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 8, no. 4, pp. 689–698, Apr. 2018.

[4] X. Tao, W. Ma, Z. Lu, and Z. Hou (2021), "Conductive particle detection for chip on glass using convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[5] L. Ruff *et al.*, "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.

[6] T. Ehret, A. Davy, J.-M. Morel, and M. Delbracio, "Image anomalies: A review and synthesis of detection methods," *J. Math. Imag. Vis.*, vol. 61, no. 5, pp. 710–743, Jun. 2019.

[7] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.

[8] B. Mohammadi, M. Fathy, and M. Sabokrou, "Image/video deep anomaly detection: A survey," 2021, *arXiv:2103.01739*.

[9] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, Mar. 2021.

[10] J. Yang, R. Xu, Z. Qi, and Y. Shi, "Visual anomaly detection for images: A systematic survey," *Proc. Comput. Sci.*, vol. 199, pp. 471–478, Jan. 2022.

[11] X. Xia *et al.*, "GAN-based anomaly detection: A review," *Neurocomputing*, vol. 493, pp. 497–535, Jul. 2022.

[12] M. E. Tschuchnig and M. Gadermayr, "Anomaly detection in medical imaging—A mini review," 2021, *arXiv:2108.11986*.

[13] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image Vis. Comput.*, vol. 106, Feb. 2021, Art. no. 104078.

[14] D. Carrera, G. Boracchi, A. Foi, and B. Wohlberg, "Detecting anomalous structures by convolutional sparse models," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–8.

[15] S. Vaikundam, T.-Y. Hung, and L. T. Chia, "Anomaly region detection and localization in metal surface inspection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 759–763.

[16] D. Carrera, F. Manganini, G. Boracchi, and E. Lanzarone, "Defect detection in SEM images of nanofibrous materials," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 551–561, Apr. 2017, doi: 10.1109/TII.2016.2641472.

[17] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, Jun. 2017, pp. 146–157.

[18] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," 2018, *arXiv:1807.02011*.

[19] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by CNN-based self-similarity," *Sensors*, vol. 18, no. 2, p. 209, Jan. 2018.

[20] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9592–9600.

[21] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," 2019, *arXiv:1911.02357*.

[22] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[23] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," 2020, *arXiv:2011.08785*.

[24] J. Yi and S. Yoon, "Patch SVDD: Patch-level SVDD for anomaly detection and segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 375–390.

[25] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," 2020, *arXiv:2005.02357*.

[26] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," 2020, *arXiv:2011.11108*.

[27] J. Pirnay and K. Chai, "Inpainting transformer for anomaly detection," 2021, *arXiv:2104.13897*.

[28] J. Yu *et al.*, "FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows," 2021, *arXiv:2111.07677*.

[29] V. Zavrtanik, M. Kristan, and D. Skočaj, "DRAEM—A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8330–8339.

[30] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," 2021, *arXiv:2106.08265*.

[31] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, Sep. 2018, pp. 161–169.

[32] S. Youkachen, M. Ruchanurucks, T. Phatrapomnant, and H. Kaneko, "Defect segmentation of hot-rolled steel strip surface by using convolutional auto-encoder and conventional image processing," in *Proc. 10th Int. Conf. Inf. Commun. Technol. Embedded Syst. (IC-ICTES)*, Mar. 2019, pp. 1–5.

[33] G. Kang, S. Gao, L. Yu, and D. Zhang, "Deep architecture for high-speed railway insulator surface defect detection: Denoising autoencoder with multitask learning," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 8, pp. 2679–2690, Aug. 2019.

[34] J. K. Chow, Z. Su, J. Wu, P. S. Tan, X. Mao, and Y. H. Wang, "Anomaly detection of defects on concrete structures with the convolutional autoencoder," *Adv. Eng. Informat.*, vol. 45, Aug. 2020, Art. no. 101105.

[35] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip connected and adversarially trained encoder–decoder anomaly detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[36] A.-S. Collin and C. De Vleeschouwer, "Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7915–7922.

[37] S. Mei, H. Yang, and Z. Yin, "An unsupervised-learning-based approach for automated defect inspection on textured surfaces," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 8, pp. 1266–1277, Jun. 2018.

[38] H. Yang, Y. Chen, K. Song, and Z. Yin, "Multiscale feature-clustering-based fully convolutional autoencoder for fast accurate visual inspection of texture surface defects," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 3, pp. 1450–1467, Jul. 2019.

[39] P. Mishra, C. Piciarelli, and G. L. Foresti, "Image anomaly detection by aggregating deep pyramidal representations," 2020, *arXiv:2011.06288*.

[40] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[41] D. S. Tan, Y.-C. Chen, T. P.-C. Chen, and W.-C. Chen, "TrustMAE: A noise-resilient defect classification framework using memory-augmented auto-encoders with trust regions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 276–285.

[42] T. Niu, B. Li, W. Li, Y. Qiu, and S. Niu, "Positive-sample-based surface defect detection using memory-augmented adversarial autoencoders," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 1, pp. 46–57, Feb. 2022.

[43] J. Hou, Y. Zhang, Q. Zhong, D. Xie, S. Pu, and H. Zhou, "Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 8791–8800.

[44] Y. Yang, S. Xiang, and R. Zhang, "Improving unsupervised anomaly localization by applying multi-scale memories to autoencoders," 2020, *arXiv:2012.11113*.

[45] X. Gui, D. Wu, Y. Chang, and S. Fan, "Constrained adaptive projection with pretrained features for anomaly detection," 2021, *arXiv:2112.02597*.

[46] Y. Liao, A. Bartler, and B. Yang, "Anomaly detection based on selection and weighting in latent space," 2021, *arXiv:2103.04662*.

[47] L. Wang, D. Zhang, J. Guo, and Y. Han, "Image anomaly detection using normal data only by latent space resampling," *Appl. Sci.*, vol. 10, no. 23, p. 8660, Dec. 2020.

[48] H. Yang, Q. Zhou, K. Song, and Z. Yin, "An anomaly feature-editing-based adversarial network for texture defect visual inspection," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2220–2230, Mar. 2021.

[49] Y. Liu, C. Zhuang, and F. Lu, "Unsupervised two-stage anomaly detection," 2021, *arXiv:2103.11671*.

[50] C.-C. Tsai, T.-H. Wu, and S.-H. Lai, "Multi-scale patch-based representation learning for image anomaly detection and segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3992–4000.

[51] Y. Chen, Y. Tian, G. Pang, and G. Carneiro, "Unsupervised anomaly detection and localisation with multi-scale interpolated Gaussian descriptors," 2021, *arXiv:2101.10043*.

[52] M. Niu, Y. Wang, K. Song, Q. Wang, Y. Zhao, and Y. Yan, "An adaptive pyramid graph and variation residual-based anomaly detection network for rail surface defects," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[53] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107706.

[54] M. Nakanishi, K. Sato, and H. Terada, "Anomaly detection by autoencoder based on weighted frequency domain loss," 2021, *arXiv:2105.10214*.

[55] T. Matsubara, K. Sato, K. Hama, R. Tachibana, and K. Uehara, "Deep generative model using unregularized score for anomaly detection with heterogeneous complexity," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5161–5173, Jun. 2022.

[56] N. Kozamernik and D. Bračun, "Visual inspection system for anomaly detection on KTL coatings using variational autoencoders," *Proc. CIRP*, vol. 93, pp. 1558–1563, Jan. 2020.

[57] W. Liu *et al.*, "Towards visually explaining variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8642–8651.

[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[59] S. Venkataramanan, K. C. Peng, R. V. Singh, and A. Mahalanobis, "Attention guided anomaly localization in images," in *Proc. Eur. Conf. Comput. Vis.* Springer, Cham, Aug. 2020, pp. 485–503.

[60] D. Zimmerer, J. Petersen, S. A. A. Kohl, and K. H. Maier-Hein, "A case for the score: Identifying image anomalies using variational autoencoder gradients," 2019, *arXiv:1912.00003*.

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

[62] Y.-T.-K. Lai and J.-S. Hu, "A texture generation approach for detection of novel surface defects," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 4357–4362.

[63] Y. T. K. Lai, J. S. Hu, Y. H. Tsai, and W. Y. Chiu, "Industrial anomaly detection and one-class classification using generative adversarial networks," in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jul. 2018, pp. 1444–1449.

[64] J. Balzategui, L. Eciolaza, and D. Maestro-Watson, "Anomaly detection and automatic labeling for solar cell quality inspection based on generative adversarial network," 2021, *arXiv:2103.03518*.

[65] Z. Zhao, B. Li, R. Dong, and P. Zhao, "A surface defect detection method based on positive samples," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, Aug. 2018, pp. 473–481.

[66] K. Komoto, S. Nakatsuka, H. Aizawa, K. Kato, H. Kobayashi, and K. Banno, "A performance evaluation of defect detection by using denoising AutoEncoder generative adversarial networks," in *Proc. Int. Workshop Adv. Image Technol. (IWAIT)*, Jan. 2018, pp. 1–4.

[67] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.* Springer, Cham, Dec. 2018, pp. 622–637.

[68] L. Zhu, Q. Zhang, and W. Wang, "Residual attention dual autoencoder for anomaly detection and localization in cigarette packaging," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2020, pp. 475–480.

[69] J. Liu, K. Song, M. Feng, Y. Yan, Z. Tu, and L. Zhu, "Semi-supervised anomaly detection with dual prototypes autoencoder for industrial surface inspection," *Opt. Lasers Eng.*, vol. 136, Jan. 2021, Art. no. 106324.

[70] J. Wang, G. Yi, S. Zhang, and Y. Wang, "An unsupervised generative adversarial network-based method for defect inspection of texture surfaces," *Appl. Sci.*, vol. 11, no. 1, p. 283, Dec. 2020.

[71] D. Zhang, S. Gao, L. Yu, G. Kang, X. Wei, and D. Zhan, "DefGAN: Defect detection GANs with latent space pitting for high-speed railway insulator," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[72] J. Li, X. Xu, L. Gao, Z. Wang, and J. Shao, "Cognitive visual anomaly detection with constrained latent representations for industrial inspection robot," *Appl. Soft Comput.*, vol. 95, Oct. 2020, Art. no. 106539.

[73] J. Yu, D. Y. Kim, Y. Lee, and M. Jeon, "Unsupervised pixel-level road defect detection via adversarial image-to-frequency transform," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1708–1713.

[74] S. Niu, B. Li, X. Wang, and H. Lin, "Defect image sample generation with GAN for improving defect recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 3, pp. 1611–1622, Jul. 2020.

[75] T. Hoshi, Y. Baba, and G. Gavai, "Railway anomaly detection model using synthetic defect images generated by CycleGAN," 2021, *arXiv:2102.12595*.

[76] O. Rippel, M. M'uller, and D. Merhof, "GAN-based defect synthesis for anomaly detection in fabrics," in *Proc. 25th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Sep. 2020, pp. 534–540.

[77] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 1530–1538.

[78] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but DifferNet: Semi-supervised defect detection with normalizing flows," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1907–1916.

[79] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 98–107.

[80] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1088–1097.

[81] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," in *Proc. BMVC* 2021, pp. 1–14.

[82] H. Chung, J. Park, J. Keum, H. Ki, and S. Kang, "Unsupervised anomaly detection using style distillation," *IEEE Access*, vol. 8, pp. 221494–221502, 2020.

[83] D. Dehaene and P. Eline, "Anomaly localization by modeling perceptual features," 2020, *arXiv:2008.05369*.

[84] Y. Shi, J. Yang, and Z. Qi, "Unsupervised anomaly segmentation via deep feature reconstruction," *Neurocomputing*, vol. 424, pp. 9–22, Feb. 2021.

[85] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "VT-ADL: A vision transformer network for image anomaly detection and localization," 2021, *arXiv:2104.10036*.

[86] R. Imamura, K. Azuma, A. Hanamoto, and A. Kanemura, "MLF-SC: Incorporating multi-layer features to sparse coding for anomaly detection," 2021, *arXiv:2104.04289*.

[87] J.-H. Kim, D.-H. Kim, S. Yi, and T. Lee, "Semi-orthogonal embedding for efficient unsupervised anomaly segmentation," 2021, *arXiv:2105.14737*.

[88] N. Li, K. Jiang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Anomaly detection via self-organizing map," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 974–978.

[89] O. Rippel, A. Chavan, C. Lei, and D. Merhof, "Transfer learning Gaussian anomaly detection by fine-tuning representations," 2021, *arXiv:2108.04116*.

[90] Y. Yan, D. Wang, G. Zhou, and Q. Chen, "Unsupervised anomaly segmentation via multilevel image reconstruction and adaptive attention-level transition," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[91] M. Tailanian, P. Muse, and A. Pardo, "A multi-scale a contrario method for unsupervised image anomaly detection," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 179–184.

[92] Y. Zheng, X. Wang, R. Deng, T. Bao, R. Zhao, and L. Wu, "Focus your distribution: Coarse-to-fine non-contrastive learning for anomaly detection and localization," 2021, *arXiv:2110.04538*.

[93] T. Nakazawa and D. V. Kulkarni, "Anomaly detection and segmentation for wafer defect patterns using deep convolutional encoder–decoder neural network architectures in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 2, pp. 250–256, May 2019.

[94] T. Tayeh, S. Aburakhia, R. Myers, and A. Shami, "Distance-based anomaly detection for industrial surfaces using triplet networks," in *Proc. 11th IEEE Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Nov. 2020, pp. 372–377.

[95] Z. Li *et al.*, "Superpixel masking and inpainting for self-supervised anomaly detection," in *Proc. BMVC*, Jan. 2020, pp. 1–12.

[96] X. Yan, H. Zhang, X. Xu, X. Hu, and P. A. Heng, "Learning semantic context from normal samples for unsupervised anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, May 2021, pp. 3110–3118.

[97] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9664–9674.

[98] J. Song, K. Kong, Y.-I. Park, S.-G. Kim, and S.-J. Kang, "AnoSeg: Anomaly segmentation network using self-supervised learning," 2021, *arXiv:2110.03396*.

[99] H. M. Schlüter, J. Tan, B. Hou, and B. Kainz, "Natural synthetic anomalies for self-supervised anomaly detection and localization," 2021, *arXiv:2109.15222*.

[100] M. Haselmann and D. P. Gruber, "Pixel-wise defect detection by CNNs without manually labeled training data," *Appl. Artif. Intell.*, vol. 33, no. 6, pp. 548–566, May 2019.

[101] N.-C. Ristea *et al.*, "Self-supervised predictive convolutional attentive block for anomaly detection," 2021, *arXiv:2111.09099*.

[102] F. Ye, C. Huang, J. Cao, M. Li, Y. Zhang, and C. Lu, "Attribute restoration framework for anomaly detection," *IEEE Trans. Multimedia*, vol. 24, pp. 116–127, 2022.

[103] T. Ulutas, M. A. N. Oz, M. Mercimek, and O. T. Kaymakci, "Split-brain autoencoder approach for surface defect detection," in *Proc. Int. Conf. Electr., Commun., Comput. Eng. (ICECCE)*, Jun. 2020, pp. 1–5.

[104] P. de Haan and S. Löwe, "Contrastive predictive coding for anomaly detection," 2021, *arXiv:2107.07820*.

[105] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[106] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[107] S. Yoa, S. Lee, C. Kim, and H. J. Kim, "Self-supervised learning for anomaly detection with dynamic local augmentation," *IEEE Access*, vol. 9, pp. 147201–147211, 2021.

[108] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, Nov. 2020, pp. 1597–1607.

[109] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[110] L. Ruff *et al.*, "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2018, pp. 4393–4402.

[111] K. Liu, A. Li, X. Wen, H. Chen, and P. Yang, "Steel surface defect detection using GAN and one-class classifier," in *Proc. 25th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2019, pp. 1–6.

[112] Y. Wang *et al.*, "Unsupervised anomaly detection with compact deep features for wind turbine blade images taken by a drone," *IPSJ Trans. Comput. Vis. Appl.*, vol. 11, no. 1, pp. 1–7, Dec. 2019.

[113] C. Hu, K. Chen, and H. Shao, "A semantic-enhanced method based on deep SVDD for pixel-wise anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.

[114] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, M. Kloft, and K.-R. Müller, "Explainable deep one-class classification," 2020, *arXiv:2007.01760*.

[115] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection," *Int. J. Comput. Vis.*, vol. 129, pp. 1038–1059, Apr. 2021.

[116] C. S. C. Tsang, H. Y. T. Ngan, and G. K. H. Pang, "Fabric inspection based on the Elo rating method," *Pattern Recognit.*, vol. 51, pp. 378–394, Mar. 2016.

[117] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *J. Intell. Manuf.*, vol. 31, no. 3, pp. 759–776, 2019.

[118] J. Božič, D. Tabernik, and D. Skočaj, "Mixed supervision for surface-defect detection: From weakly to fully supervised learning," *Comput. Ind.*, vol. 129, Aug. 2021, Art. no. 103459.

[119] J. Gan, Q. Li, J. Wang, and H. Yu, "A hierarchical extractor-based visual rail surface inspection system," *IEEE Sensors J.*, vol. 17, no. 23, pp. 7935–7944, Dec. 2017.

[120] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *Vis. Comput.*, vol. 36, no. 1, pp. 85–96, 2018.

[121] K. Sohn, J. Yoon, C.-L. Li, C.-Y. Lee, and T. Pfister, "Anomaly clustering: Grouping images into coherent clusters of anomaly types," 2021, *arXiv:2112.11573*.

[122] P. Bergmann, X. Jin, D. Sattlegger, and C. Steger, "The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization," 2021, *arXiv:2112.09045*.

[123] W. H. Chu and K. M. Kitani, "Neural batch sampling with reinforcement learning for semi-supervised anomaly detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 751–766.

[124] L. Chen, Z. You, N. Zhang, J. Xi, and X. Le, "UTRAD: Anomaly detection and localization with U-transformer," *Neural Netw.*, vol. 147, pp. 53–62, Mar. 2022.

[125] S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, and U. Genc, "Anomalib: A deep learning library for anomaly detection," 2022, *arXiv:2202.08341*.

[126] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 947–969, Apr. 2022.

Xian Tao (Member, IEEE) received the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 2016.

He is currently an Associate Professor with the Research Center of Precision Sensing and Control, IACAS, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing. His research interests include deep learning and automated industrial surface inspection.

Xinyi Gong received the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 2019.

He is currently an Associate Professor with the Research Center of Precision Sensing and Control, IACAS. His research interests include computer vision, image processing, pattern recognition, and machine learning.

Xin Zhang received the Ph.D. degree from the School of Technology, Beijing Forestry University, Beijing, China, in 2017.

He has been an Associate Professor with the Beijing Technology and Business University, Beijing, since 2020. His research interests include applications in the integration of blockchain and AI.

Shaohua Yan received the B.Sc. degree in automation from Harbin Engineering University, Harbin, China, in 2019. He is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems with the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China.

His current research interests include intelligent robot control and machine vision.

Chandranath Adak (Senior Member, IEEE) received the Ph.D. degree in analytics from the University of Technology Sydney, Sydney, NSW, Australia, in 2019.

He is currently an Assistant Professor with the Indian Institute of Technology Patna, Bihar, India. His research interests include computer vision and deep learning.