IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 September 2023, accepted 9 November 2023, date of publication 13 November 2023,
date of current version 17 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3332475

RESEARCH ARTICLE

A Novel MAE-Based Self-Supervised Anomaly Detection and Localization Method

YIBO CHEN[1], HAOLONG PENG[2], LE HUANG[3], JIANMING ZHANG[1], AND WEI JIANG[1]
[1]College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
[2]Zhejiang Polytechnic Institute, Zhejiang University, Hangzhou 310027, China
[3]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China

Corresponding authors: Jianming Zhang (ncsl@zju.edu.cn) and Wei Jiang (jiangwei_zju@zju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62173302.

ABSTRACT Despite significant advancements in self-supervised anomaly detection, multi-class anomaly detection tasks still pose substantial challenges. Most existing methods require individual network training for each category of objects. This paper presents a novel end-to-end approach for multi-class anomaly detection: self-supervised Mask-pretrained Anomaly Localization Autoencoder (MALA). Firstly, the masked autoencoder (MAE) and Pseudo Label Prediction Module (PLPM) are utilized to recover and perceive normal image patterns. Subsequently, the encoder weights are frozen for further end-to-end network training to predict anomalous maps directly. Token Balance Module(TBM) facilitates anomalous perception and improves anomaly segmentation. By utilizing the Visual Transformer and employing image inpainting as a proxy task, remarkable generalization results are achieved. The proposed method demonstrates its applicability across diverse styles of industrial products. Experiments are conducted on MVTech AD, VisA, KolektorSDD2, and MT datasets, achieving state-of-the-art results in multi-task anomaly detection and segmentation tasks. Specifically, we obtain image AUROC of 98.% and pixel AUROC of 97.1% on the MVTech AD dataset, pixel AUROC of 97.1% on the VisA dataset, and pixel AUROC of 98.7% on the KolektorSDD2 dataset.

INDEX TERMS Defect localization, self-supervised learning, visual transformer (ViT), masked autoencoder (MAE), industrial products.

I. INTRODUCTION

Visual anomaly detection has been applied to various domains, including medical imaging [1], [2], transportation object or sign detection [3], [4], video surveillance [5], [6], [7], [8], [9], and industrial products [10], [11], [12], [13]. With the evolution of implication fields and the complexity of image representations, traditional methods [14], [15], [16] might not suit anomaly detection and segmentation tasks, in addition to products with typical appearance regularity [17], [18]. Thus, more effective architecture and pipelines are devised with the advent of machine learning methods and deep learning networks [19], [20], [21]. Due to the laborious and costly labeling of supervised learning tasks, unsupervised and self-supervised methods have become research hotspots in recent years, and many datasets suitable for such tasks have emerged [22], [23], [24]. Convolution neural network (CNN) [25], [26], [27], [28], memory bank [21], [29], [30], [31], and distribution assessment [32], [33], [34], [35] are three common types trained for individual datasets shown in Fig. 1 (a).

Several researchers have paid more attention to challenging multi-class training tasks [36], [37], [38], as shown in Fig.1(b). Unlike one-for-one setups, comprehensive categories put a higher standard on the model's generalization and stability. For instance, the model needs to deal with bias problems. If the parameters tilt toward specific shapes or styles of several industrial products, the model might sink into the local optimum instead of reaching an overall balance among all the types. **1**)Deep CNN-based reconstruction [37] makes the anomaly map prediction process redundant. Visual-Transformer-based (ViT) [39] models and masked autoencoders (MAE) [38], [40], [41], [42] are promising structures to be applied to multi-style anomaly localization

The associate editor coordinating the review of this manuscript and approving it for publication was Guangcun Shan.

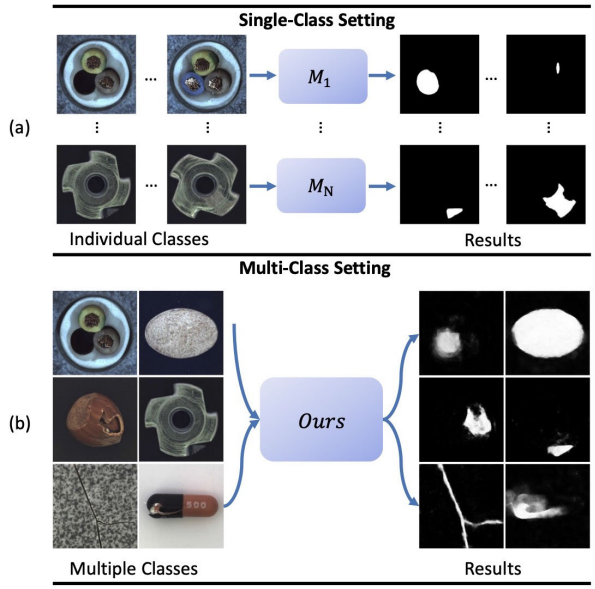
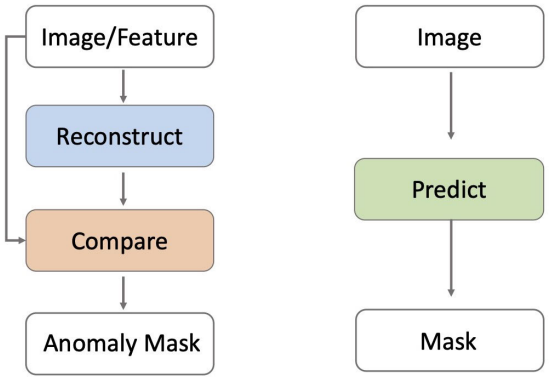


FIGURE 1. MALA is an end-to-end self-supervised mask-pretrained anomaly localization method, as shown in (b). Different from (a): Traditional individual classes detection method(N models for N classes), our method can apply to multiple classes using only one model(one model for N classes).



(a). Reconstruction-based anomaly localization.

(b). End-to-End Structure.

FIGURE 2. Flow charts under different paradigms. (a) indicates the process of Reconstruction and anomaly localization. (b) indicates the process of End-to-End.

and detection task. **2)**You et al. [36] combine feature jittering with an attention mechanism to increase the model's power, yet their model still has the potential to optimize segmentation results. On the other hand, properly trained tokenizers, extracted prototypes, and comparatively deep encoder-decoder [38] incur substantial inference costs and a barrier to typical feature preserving. Previous reconstruction-based methods are indicated in Fig.2(a).

To increase the prediction accuracy and lower the threshold for feature extraction in tough various types of training sets, we propose Mask-pretrained Anomaly Localization Autoencoder(MALA), applying a two-stage training paradigm and end-to-end defective segmentation pipeline. The end-to-end idea is shown in Fig.2(b). In the first stage shown in Fig.3 Stg.1, MALA occludes parts of normal images and trains the encoder-decoder of ViT structure to predict the whole image with the remaining. MALA contains a pseudo labels prediction module (PLPM) to train the model in a generative-adversarial way. In the second stage, shown in Fig.3 stg.2, MALA takes in artificial anomaly images and outputs defect segmentation masks. Instead of occluding suspicious areas [38] and achieving feature-level [36] or image-level reconstruction [37], MALA directly obtains the anomaly map and skips unnecessary comparison processes. To better help the anomalous perception and improve anomaly segmentation ability, we design a Token Balance Module(TBM). TBM fuses multi-scale and multi-level features and dramatically reduces the probability of pixel-level misjudgment.

In conclusion, the main contributions of this paper are as follows:

1) To cope with tough multi-class anomaly detection tasks, we propose a MAE-based framework MALA. The training process is divided into two stages. In the first stage, MALA is equipped with TBM trained in a generative-adversarial way to reconstruct normal images. In the second stage, MALA aims to predict anomaly areas of artificial defective images.

2) We propose a novel TBM module that effectively integrates multi-level feature tokens and preserves valuable information as much as possible. Meanwhile, MALA discards unnecessary comparison modules and uses an end-to-end segmentation pipeline.

3)MALA achieves competitive results on MVTech AD and VisA for both image-level and pixel-level. Moreover, MALA demonstrates its practicality and robustness against noise on the KolektorSDD2 [43] and MT [44] datasets, maintaining state-of-the-art(SOTA) performance.

II. RELATED WORK

Recent studies on frame or image-based industrial anomaly detection can be categorized based on their functional and target emphasis into the following groups: few-shot or zero-shot methods, multi-modal methods, masked image reconstruction modeling, and class-generalization methods. Generally, Few-shot or zero-shot methods [45], [46], [47], [48] are centered on extracting crucial and representative features of normal patterns to distinguish abnormal parts. In addition, pretrained vision-language models [49], [50] have drawn attention from various fields and inspired researchers to increase robustness in a new way of multi-modal thoughts [51], [52], [53] and make up for the possible neglect of details by sole vision perception.

Video-based researches [3], [4], [5], [9] about object detection or tracking thrive and improve the availability and practical application value. Liang et al. put forward Detect-Former [4] and improve sparse R-CNN with ResNeSt [3]. Both of them greatly optimize the accuracy of traffic anomaly detection. Wang et al. [5] propose a one-class abnormal event

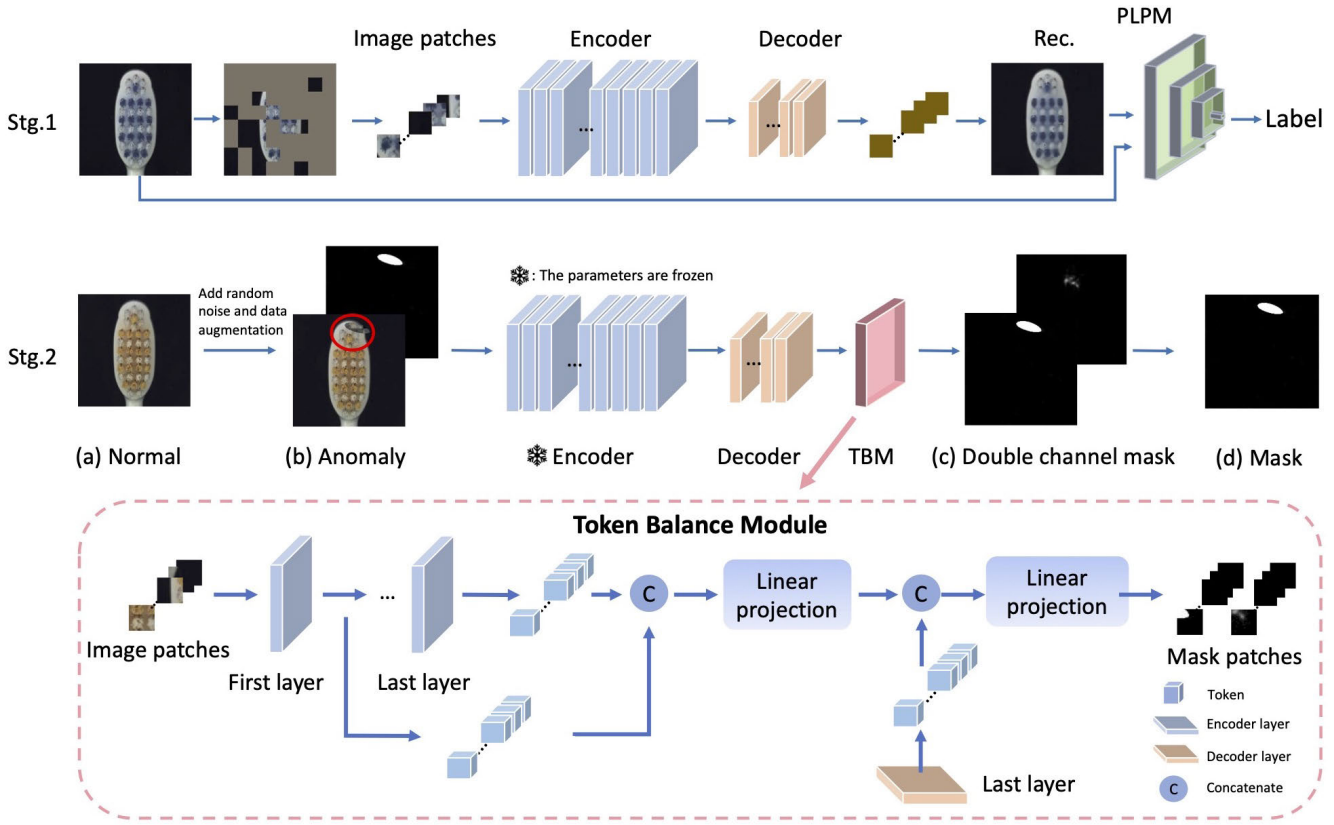


FIGURE 3. Stg.1 The overview of perceiving normal patterns of MALA in stage 1. After the picture is covered by a large portion(e.g.75%), it is sent to MAE for reconstruction, with Prediction Module(PLPM) distinguishing pseudo label types of reconstructed and original images generatively and discriminatively. Stg.2 The framework of MALA in stage 2. The model is composed of three parts:(1) Add artificial anomaly global or local Perlin Noise. (2) Frozen the encoder and activate decoder parameter settings from stage 1. (3) Balance tokens to capture both local and global details. Acquire a double-channel map and take the first channel as the output mask.

detection network (AED-Net) based on principal component analysis across various scenes. Liu et al. [9] propose a tracking algorithm based on the Siamese network, combining fuzzy inference to visual tracking with bounding boxes and anomaly maps. Methods ahead are based on multi-frame information, while the industrial anomaly detection task lacks continuous frame-level regularity and requires pixel-wise segmentation and the tolerance of different normal products in a comparatively static process.

Reconstruction-based methods are often self-supervised and combined with anomaly simulation [26], [46], [54], [55], [56], [57]. Defect-GAN [46] is one of the earliest research that proposes to inject controllable synthetic defects and random noise to train images. DRA [58] disentangles the simulated anomalies into seen, pseudo and latent-residual types. JNLD [55] proposes a defective simulation strategy based on just noticeable distortion. Nevertheless, researchers add anomalies whose texture and shapes might not be similar to real ones, aiming to help the model fully be aware of normal patterns [46], [56], [57], [59]. To enhance the randomness of the noise, they apply Perlin-Noise and other irregular shapes to make both the generation and outcome unexpected. Similarly, JDRSS [35] injects internal and external factor-induced anomalies to enhance the sensitivity to novelties. In addition to adding noise or anomalies to images, studies have tried to augment extracted features [36], [60].

Masked image prediction is another way to perform self-supervised learning. Unlike anomalous simulation techniques, these methods [38], [40], [41], [61] try to fulfill whole images with clues from unmasked parts. Early methods [41], [62] are proposed in other fields. MAE [41] trains the model to pad the missing pixels for the latter classification, and MaskFeat [62] aims to assess the features of the masked areas. Zhang et al. [63] pretrain an inpainting GAN with global and local generators to infer masked areas and later procedures combined with periodical noise injection to enhance data. Apart from image-level masking, self-supervised predictive convolutional attentive block [64] can also occlude features in arbitrary layers. Besides, the flexible blocks can be easily added to CNN or Transformer backbones and proved effective in many fields.

Visual transformer(ViT), apart from MAE-based architecture, is a reliable pipeline to be a promising deep learning

baseline for anomaly segmentation. CrackFormer [65] uses self-attention blocks to extract cross-channel context and detect fissure defects. DefT [54] uses ViT and CNN to extract global and local tokens individually. Szarski et al. [12] combine ViT with NF to generate probabilistic anomaly maps. ViTALnet [66] adopts ViT as a feature extractor and captures global semantics for downstream detection. RDAD [67] introduces a channel transformer to fuse features and increase sematic compatibility in the encoder-decoder framework.

Generalization performance on various datasets of industrial or medical fields [50], [58], [64] has been proved in various studies. Lee et al. [68] use limited annotated images to adjust the meta-learning-based structure and apply the model trained on several classes to cross-domain sets' defect discrimination. GLCF [50], containing multi-scale patch embedding and semantic aggregation, applies semantic bottleneck implemented by Vit. GLCF can discriminate logical and appearance anomalies if pretrained for a specific industrial or medical category. Jeong et al. [69] propose the few-shot window-based CLIP (WinCLIP) combining words with templates and aligning text with multi-scale image features extraction.

Parameter adaptation, namely adaptation to various types or classes, is the focus of this paper. Multi-class anomaly detection studies [36], [37], [38] research multi-class adaptive capacity. You et al. [36] apply a neighbor-masked attention module with feature jittering and a layer-wise query decoder to obtain reconstructed tokens left for anomaly detection. OmniAL [37] injects just noticeable distortion to normal images and adopts a unified CNN-based network with DiffNeck to locate anomaly regions. Similar to MAE, PMAD [38] pretrains a ViT-based encoder-decoder to recover the masked images. During inference, PMAD occludes suspicious patches of artificial defect images and passes through the encoder-decoder to determine uncertainty regions subsequently.

As introduced above, most methods aim to tune pipelines on a specific class. They sacrifice more parameters for training and inference costs due to a separate model for each class. Besides, the class-to-class fine-tune results might still fall behind the multi-class methods due to a lack of cross-type perception and conclusion. Models for the multi-class task are more sensitive to the details of different types and learn more discriminative cues during the training. In Section IV of this paper, we further discuss and display various indexes and comparisons.

III. APPROACH

We propose a novel end-to-end self-supervised training paradigm MALA based on MAE. The aim is to enable anomaly detection and localization across diverse industrial products. We design different training stages that focus on distinct pretext tasks and adapt flexible and adjustable pipelines, specifically the anomaly synthesis of DRAEM [26] and the baseline of robust inductive MAE, leveraging the global and positional perception capabilities of ViT.

In the first stage, the network comprises an encoder, a decoder, and a Pseudo Label Prediction Module (PLPM). This stage aims to generate complete images from masked images with TBM improving authority. In the second stage, the structure and parameters of the encoder and decoder from the first stage are retained. MALA contains a Token Balance Module(TBM) to fuse multi-scale and multi-level tokens. Due to well-trained modules in the second stage, the network can predict pixel-level anomaly maps and image-level anomaly scores among diverse products and various styles of actual defects.

A. STAGE 1: PERCEIVE NORMAL PATTERNS

In this stage, the objective is to perceive all types of details related to various industrial products. MAE serves as a promising baseline and reconstructs considerable images even when a large portion (e.g. 75%) is randomly covered in each epoch. However, challenges arise when dealing with positional or appearance variations in industrial products for tolerable minor changes.

To address the problem above, a more adaptive structure is proposed. Assume a training batch of a specific product, denoted as $X = \{x_i \mid i \in 1, \ldots, n\}$, where n is the number of randomly selected samples from the normal dataset. As shown in Fig.3 Stg.1, a single image x_i is masked by a fixed ratio in each epoch. The positional information combined with uncovered parts is collected and rearranged, denoted as $\hat{x}_i$. All the scratched permutations in the batch are $\hat{X} = \{\hat{x}_i \mid i \in 1, \ldots, n\}$. The encoder, denoted as φ_{enc}, uses deep ViT-based blocks. After passing through φ_{enc}, the normal class token and feature tokens are extracted, denoted as $Z = \{z_{cls}, z_i \mid i \in 1, \ldots, j\}$.

$$z_i = \varphi_{\text{enc}}(\hat{x}_i), \quad i = 1, \ldots, j \tag{1}$$

Since the goal is not classification, the class token z_{cls} is discarded in this structure.

Utilizing positional encoding, the tokens of masked areas are located and padded with zeros. The padded token set for a batch is represented as $\hat{Z} = \{\hat{z}_i \mid i \in 1, \ldots, n\}$. $\hat{Z}$ then passes through decoder φ_{enc} consisting of attention blocks. Inspired by previous research, we choose a smaller and shallower ViT-based decoder than the encoder. The decoder output $\bar{z}_i$ can be formulated as:

$$\bar{z}_i = \varphi_{dec}(\hat{z}_i) \quad i = 1, \ldots, j \tag{2}$$

Pseudo Label Prediction Module (PLPM) is a self-supervised learning method. Considering that defect detection is mostly small targets, although MAE's reconstruction effect is already excellent, due to the strictness of defect detection in industrial scenarios, we still need to enhance the power of feature perception and extraction. To this end, we introduced the self-supervised learning method PLPM to perform comparative predictions on true and false labels to

enhance the model's learning and representation capabilities for the details of the object to be detected.

The linearly projected and rearranged normal images are predicted, denoted as $\bar{X} = \{\bar{x}_i \mid i \in 1, \ldots, n\}$. Subsequently, the pseudo labels are assigned 0 for the original images and 1 for the reconstructed ones. Pseudo Label Prediction Module, consisting of convolutional layers, is trained to distinguish label types. This concept is akin to the idea behind generative adversarial networks. The predicted label vectors are lb_o and lb_p for original and reconstructed images. The introduction of PLPM significantly proved its ability in subsequent ablation experiments.

The training loss is a combination of three components: the L1 loss, the cross-image structure similarity index measure (SSIM) [70], and the cross-entropy loss (CE) of pseudo labels. The normalized L1 loss is calculated as the multi-channel pixel-wise absolute average error between the reconstructed and original images. Meanwhile, the SSIM measures the structural similarity between them. Cross-entropy loss is computed between the predicted labels and the assigned labels.

$$\begin{cases} L_1(X, \bar{X}) = \dfrac{1}{n}\sum_{i=1}^{n} \text{avg}\,(\text{abs}\,(x_i - \bar{x}_i)) \\ L_{SSIM}(X, \bar{X}) = \dfrac{1}{n}\sum_{i=1}^{n}[1 - \text{SSIM}(x_i, x_i)] \\ L_{CE}(lb_o, lb_p) = -\dfrac{1}{2n}\sum_{i=1}^{n}\log[1 - \sigma(lb_{o,i})] \\ \qquad\qquad -\dfrac{1}{2n}\sum_{i=1}^{n}\log[1 - \sigma(lb_{p,i})] \\ L_{stg1} = \alpha_1 L_1(\bar{X}, \hat{X}) + \alpha_2 L_{SSIM}(\bar{X}, \hat{X}) \\ \qquad\qquad +\alpha_3 L_{CE}(lb_o, lb_p) \end{cases} \tag{3}$$

where $\alpha_1, \alpha_2, \alpha_3$ are the balancing hyper-parameters.

B. STAGE 2: PREDICT ARTIFICIAL ANOMALY MAPS

In this stage, we transform the previous network into an end-to-end anomaly-assessment pipeline. MALA randomly takes m examples from each normal object set, denoted as $X = \{x_i \mid i \in 1, \ldots, m\}$. The artificial anomaly types contain global Perlin Noise(PN), as mentioned in DRAEM [26], and local PN acquired from foreground segmentation. The encoder is frozen while the decoder remains adaptive. The challenges are as follows: **1)** Even if tokens from the encoder contain anomalies, the encoder needs to ensure that the encoded output is informative enough to recognize anomaly areas. Indeed, there is still a gap between the recovery pretext task in Stage 1 and defect segmentation in the second stage. **2)** Additionally, the decoder must overcome the differences between synthetic and real defects. **3)** Furthermore, given the abundance of object types, it is challenging to achieve multi-class and multi-style balance.

To cope with the problems above, we adopt an end-to-end decoder structure with TBM, as depicted in Fig.3. We continue to improve the model's ability to learn details for higher productivity in industrial production. Since the defects of small objects are subtle, we need more receptive fields, so we propose the TBM module to fuse multi-scale and multi-level features. Feature fusion of different scale dimensions significantly improves the feature learning ability of the model. The primary objective is to capture local and global details, encompassing texture and semantic knowledge. However, local features are still important to obtain local geometric details, especially for small objects. Motivated by this, we add a Linear projection to realize adaptive fusion.

To be more specific, consider the input anomaly samples' set as $\bar{X} = \{\bar{x}_i \mid i \in 1, \ldots, k\}$ and anomaly masks' set as $\bar{Y} = \{\bar{y}_i \mid i \in 1, \ldots, k\}$. $\bar{X}$ passes through the encoder and is transformed into corresponding tokens $\bar{Z} = \{\bar{z}_i \mid i \in 1, \ldots, k\}$. The decoder φ_{dec2} is trained to approximate anomaly maps.

$$\bar{z}_i = \varphi_{dec2}(\hat{z}_i), \quad i = 1, \ldots, k \tag{4}$$

The details of the Token Balance Module(TBM) are shown in Fig.3. The output tokens from the first and the last layers of the encoder are concatenated and linearly projected to the dimension of the decoder. Then the transformed tokens and the output of decoder's last layer are concatenated and linearly projected as the final two-channel defect maps, denoted as $\hat{Y}$. The two-layer output enhances the robustness and stability of multi-level feature token fusion. The first channels of $\hat{Y}$, denoted as $\hat{Y}[0]$, are the eventual defective segmentation masks. The calculation of TBM is shown in eq.(5).

$$\hat{Y} = LP(Con(LP(Con(\hat{Z}_{enc,first}, \hat{Z}_{enc,last}))), \bar{Z}_{dec,last}) \tag{5}$$

In the TBM, the *LP* indicates a linear projection layer and *Con* indicates a concatenation operator.$\hat{Z}_{enc,first}$ and $\hat{Z}_{enc,last}$ are the output of the first and the last layer of the encoder individually. And $\bar{Z}_{dec,last}$ is the output set of the decoder's last layer.

In this stage, the training loss is combined with L1 loss L_1, focal loss L_{focal} [71] to calibrate predicted results. The focal loss helps precisely localize relatively less prominent anomaly regions. The stage2 loss L_{stg2} can be defined as:

$$L_{stg2} = \beta_1 L_1\left(\bar{Y}, \hat{Y}[0]\right) + \beta_2 L_{focal}(\bar{Y}, \hat{Y}) \tag{6}$$

In this procedure, reconstruction is discarded. On the one hand, it is necessary to prevent the pipeline from becoming overly reliant on detecting conspicuous color anomalies while potentially ignoring the importance of accurately reconstructing the normal patterns. On the other hand, reliably predicted masks are obtained from comprehensive perception. To facilitate understanding the whole process, the two training stages above are shown as pseudo-code Algorithm 1 and Algorithm 2.

C. STAGE 3: VALIDATE THE REAL ANOMALY IMAGES

In this stage, we freeze all the parameters of the pipeline and aim to obtain the predicted anomaly masks

Algorithm 1 Training Algorithm of Stage 1

Input: A set of normal images $X = \{x_i \mid i \in 1, \ldots, n\}$
Output: A set of predicted full normal images $\bar{X} = \{\bar{x}_i \mid i \in 1, \ldots, k\}$

1: Obtain a set of scratched permutations $\hat{X}$ by randomly masking the samples of X. Pass $\hat{X}$ through the encoder subsequently as following
2: $z_i = \varphi_{\text{enc}}(\hat{x}_i), \quad i = 1, \ldots, j$
3: The token set is padded for a batch, represented as $\hat{Z}$ and passes through the decoder.
4: $\bar{z}_i = \varphi_{dec}(\hat{z}_i) \quad i = 1, \ldots, j$
5: The set is linearly projected and arranged as $\bar{X}$. The eventual loss between $\hat{X}$ and $\bar{X}$ concludes L_1, L_{SSIM} and L_{CE} represented as eq. (3).

$\bar{Y} = \{\bar{y}_i \mid i \in 1, \ldots, k\}$. However, relying solely on extreme pixel predictions and ignoring the overall image comparison may not be the most optimal approach. To further gain image-level anomaly assessment, the pipeline incorporates a new filter convolution layer to feedback objective scores.

$$S_{anomaly} = max(\bar{Y}[0] * f_{k \times k}) \tag{7}$$

The image-level scoring vector of anomaly evaluation, denoted as $S_{anomaly}$, is generated by making full use of the predicted anomaly masks $\bar{Y}$. The $f_{k \times k}$ indicates the average pooling kernel size is $k \times k$ and $*$ means the convolution operator.

IV. EXPERIMENT RESULTS

A. DATASETS

We mainly measure the training paradigm and pipeline structure with two widely used challenging public datasets, MVTec Anomaly Detection (MVTec AD) [22] and Visual Anomaly Dataset(VisA) [24]. In addition, to further validate the practicality of MALA, we also train and test on KolektorSDD2 [43] and MT [44] datasets.

MVTec AD dataset contains 15 diverse industrial product types, including 5 texture classes and 10 object classes. The dataset comprises 3,629 normal training images and 1,725 normal or abnormal testing images. Additionally, there are various types of defects for individual products, such as 'bent' or 'scratch' in the case of hazelnuts, adding up to over 70 different defect types in total. All the anomaly images are appended with pixel-wise labeled masks denoting the segmentation of unusual areas. The usual testing masks of examples can be generated conveniently of zero matrices in the same size with input images. The input images in the dataset have varying dimensions, ranging from 700×700 to 1024×1024 pixels, and can have either 1 or 3 color channels. Anomalies with diverse sizes and shapes within the dataset set a higher standard for discrimination, requiring anomaly detection methods to accurately identify and localize anomalies of different scales and forms.

Algorithm 2 Training Algorithm of Stage 2

Input: A set of normal images $X = \{x_i \mid i \in 1, \ldots, n\}$
Output: A set of self-supervised defective segmentation masks $\hat{Y}[0] = \{\hat{y}[0]_i \mid i \in 1, \ldots, k\}$

1: Add PN into X and obtain a set of artificial anomaly samples $\bar{X}$ and anomaly masks' set $\bar{Y}$. $\bar{X}$ passes through the frozen encoder and is transformed into corresponding tokens $\bar{Z}$
2: $\hat{z}_i = \varphi_{\text{enc}}(\bar{x}_i), \quad i = 1, \ldots, j$
3: $\bar{z}_i = \varphi_{dec2}(\hat{z}_i), \quad i = 1, \ldots, k$
4: The TBM fuses the multi-level feature tokens and outputs the eventual two-channel defect maps, denoted as $\hat{Y}$.
5: $\hat{Y} = LP(Con(LP(Con(\hat{Z}_{enc,first}, \hat{Z}_{enc,last}))), \bar{Z}_{dec,last})$ In TBM, the *LP* indicates the linear projection layer, and *Con* indicates the concatenation operator. The $\hat{Z}_{enc,first}$ and $\hat{Z}_{enc,last}$ are the output of the first and the last layer of the encoder individually. And $\bar{Z}_{dec,last}$ is the output of the decoder's last layer. The first channels of $\hat{Y}$, denoted as $\hat{Y}[0]$, are the eventual defective segmentation masks. The loss is listed as eq. (6).

VisA dataset contains 12 different industrial product types with 10,821 high-resolution images, including multiple instances or a single object. There are 1,200 defective samples for testing in aggregation. Nevertheless, several single-object classes, like printed circuit boards (PCB) containing details about traces, components, and other intricate features, are complicated. Although the anomalous types are not labeled as clearly as MVTec AD, VisA still includes various logical flaws like missing modules and surface flaws like colored stains or cracks. Noticing the anomaly masks' values are constrained from 0 to 1, and it is of necessity to scale in the proper range to ensure accurate analysis and interpretation of the anomalies.

KolektorSDD2 dataset consists of surface defects that vary in shape and color, including small and large-scale ones. This dataset includes over 2000 defect-free training samples.

MT Defect dataset contains anomalous magnetic tile images from real industrial scenarios. There are five types of defects, including blow-hole, break, uneven, fray, and crack. We still use only 100 from over 1900 normal samples in this dataset.

The wide adaptability and generalization of our approach set a baseline for addressing the demands of complex industrial scenarios.

B. METRICS

The Area Under the Receiver Operating Characteristic Curve (AUROC) is a commonly used evaluation method in anomaly detection and localization. Image-level AUROC provides effective judgments on the performance of classification models, while pixel-level AUROC points out the quality

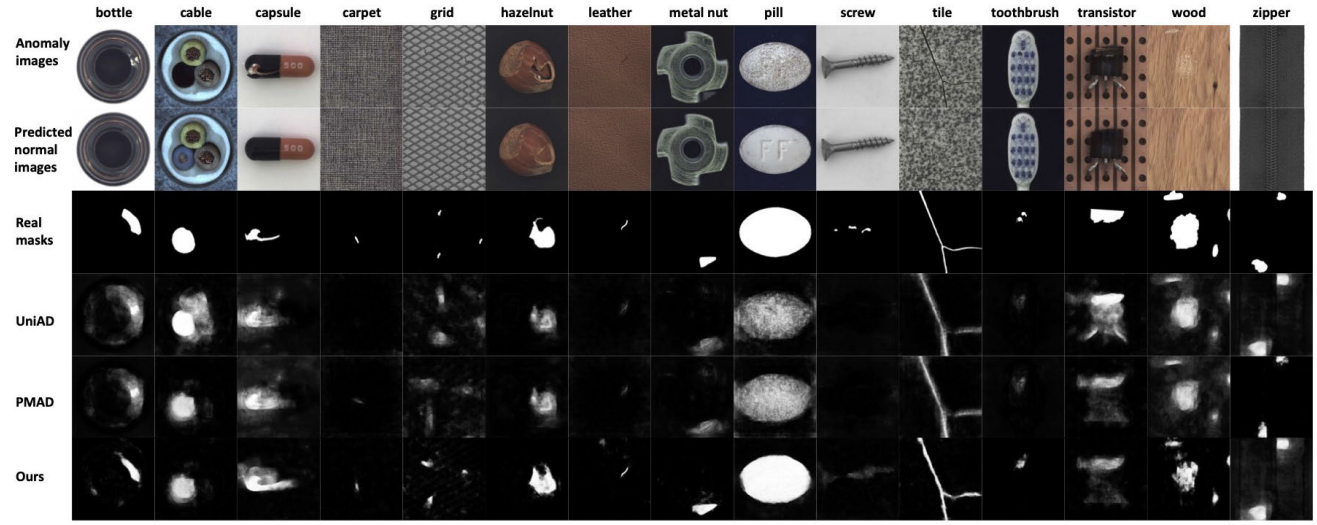


FIGURE 4. Visualization of the anomaly segmentation results on MVTech AD dataset.

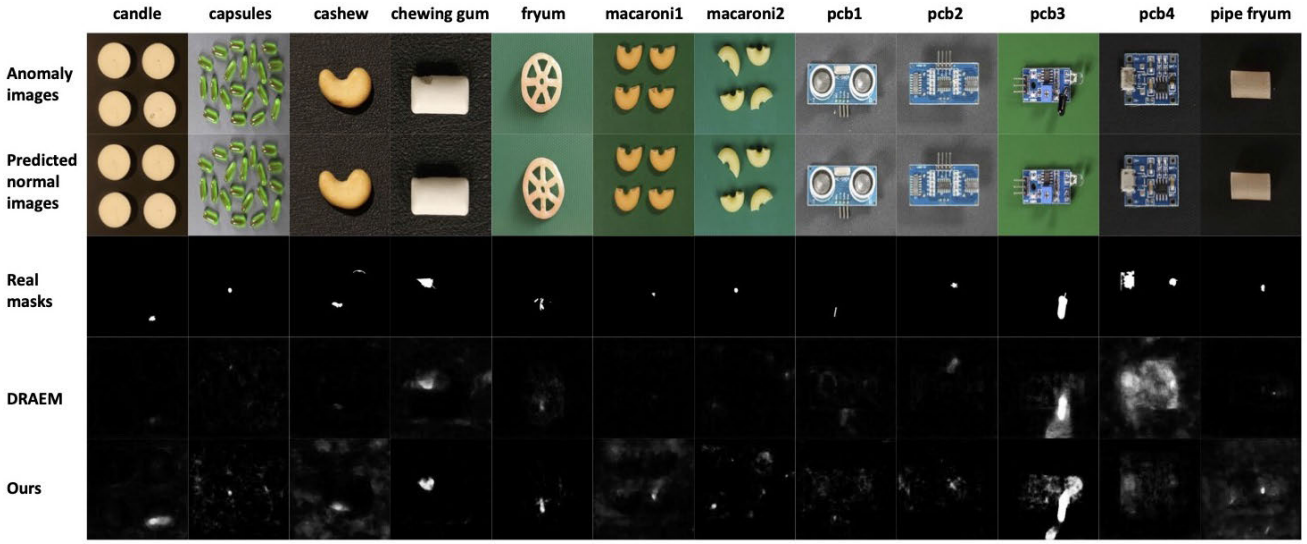


FIGURE 5. Visualization of the anomaly segmentation results on VisA dataset.

of anomaly prediction on a finer scale. A higher pixel-level AUROC score indicates a more precise prediction and segmentation of anomalies at the pixel level. The inferred anomaly masks, closely related to the image-level scores, heavily rely on the model's reconstruction capability.

In summary, both the datasets and metrics are challenging and provide convincing evidence to compare the results among multiple detection and localization techniques. The following competitive experiment results further prove the power of our architecture.

C. EXPERIMENT SETTINGS

All the training stages are completed on a single GPU (GeForce RTX 2080 Ti). We resize all input images and masks to a consistent size of 224 × 224 pixels. We use a 3-channel representation for the images, while the masks are represented as single-channel images. Our encoder employs the standard large-ViT preparation technique, including sine-cosine positional encoding and 14 × 14 patch size. The large-ViT model consists of 24 ViT blocks in the encoder, where each feature token has a dimension of 1024. A single decoder comprises 8 blocks, each with a feature token dimension of 512. We add linear projection layers at the start and end of the decoder.

In stage 1, the initial parameters of the encoder and decoder are pretrained on the ImageNet1K dataset from He's team. The PLPM's parameters are initialized randomly. During training, the model is updated using the AdamW optimizer

with a weight decay of 0.05 and momentum β of (0.9, 0.95). Similar to MAE, the learning rate lr. is applied with linear scaling $lr = lr_{base} \times batch\ size\ /256$. The lr_{base} is set as 1e-3, while the batch size is 128. In this stage, the input includes 27 product categories, and 200 normal images of each category are randomly selected and masked for 75% in each epoch due to random seeds. All the multi-class images are simultaneously trained for 2000 epochs.

In stage 2, we freeze the encoder parameters. The decoder parameters are initialized from stage 1, and the parameters of TBM are randomly initialized. Unlike stage 1, we randomly select only 80 samples for training. Among these samples, 64 are injected with defects, while the others remain unchanged for each industrial item. The abnormal samples, texture, position, and shapes are randomly updated in every epoch. The defective parts originate from the Describable Textures Dataset (DTD). Similar to DRAEM, the anomaly texture augmentation methods include transparent adjustment, posterizing, solarizing, and so forth. The learning rate lr is fixed, while the *batch size* is set to 32. Both the normal and abnormal samples are fed into the pipeline trained for 700 epochs.

D. COMPARISON AND PERFORMANCE

1) PERFORMANCE ON MVTEC AD

Pixel-level and image-level AUROC unified-training results of the MVTec AD dataset are shown in Table 1. The anomaly segmentation results are shown in Fig.4. It should be mentioned that both the MVTec AD dataset and VisA dataset are input simultaneously in our configuration to be more challenging.

Compared to other results, our method has improved the anomaly detection result for the capsule class by 6.6% compared to the second-best result of 91.2% obtained by PMAD. Besides, the result of the pill also achieves a notable 2.8% increase in contrast with PMAD. The screw, capsule, and pill classes achieved 95.6%, 97.8%, and 99.3% respectively. Furthermore, average results surpass all previous methods to become state-of-the-art(SOTA). While some individual results may not be exceptionally outstanding, the average image-level AUROC has improved by 1.9% compared to the second-best result of 96.7% achieved by UniAD, and it reached the SOTA level with a result of 98.6%.

Our method also maintains competitiveness in pixel-level segmentation. The mean pixel-level AUROC has reached the SOTA with the result of 97.1%. Eight categories have also reached SOTA, including grid (99.3%), hazelnut (99.4%), leather (99.5%), pill (98.2%), tile (99.3%), toothbrush (99.4%), wood (97.9%) and zipper(99.2%). These excellent results undoubtedly demonstrate MALA's significant discriminatory potential and excellent competitiveness.

As shown in Table 1 and Table 2, even though some methods perform considerably under the single-class task, they may not be able to keep satisfying performance under the multi-class situation. That is related to their ability to discern the key patterns of various targets. The cost of the representative methods is analyzed in section IV-F.

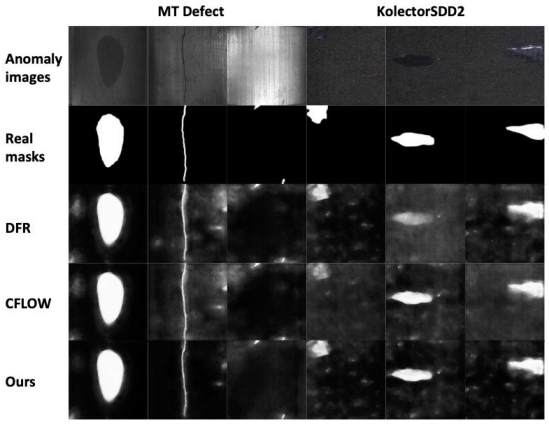


FIGURE 6. Visualization of the anomaly segmentation results on KolektorSDD2 and MT Defect datasets.

2) PERFORMANCE ON VISA

Image-level results on the VisA dataset, as presented in Table 3, the anomaly segmentation results are shown in Fig.5, demonstrate the generalization capabilities of our model and experimental setup. Despite the challenges posed by plentiful inputs, our results remain competitive and stable. On the other hand, the pixel-level results in Table 3 present even more outstanding performance, maintaining a lead of over 1.1% compared to the second-best results of 96.1%, achieving SOTA with an excellent result of 97.1%. Notably, the pixel-level indicators for categories such as fryum and pipe-fryum, respectively, achieve 99.1% and 99.7%, which shows significant improvements. Although the average image-level AUROC may not be the absolute best, on average, it still shows a competitive advantage compared to OmniAL and UniAD. MALA, with the result of 93.0%, is second only to the current SOTA model OminiAL.

3) PERFORMANCE ON KOLEKTORSDD2 AND MT DATASETS

Our method applied to industrial datasets KolektorSDD2 and MT datasets also achieves notable results, as shown in Fig.6. Unified-training results of the two datasets are shown in Table 5. Our experiment, with a result of 98.7%, surpasses the leading ViTALNet on the KolektorSDD2 dataset and reaches the state-of-the-art level. In addition, the performance on the MT Defect dataset has also achieved notable 1.1% and 5.6% increases in contrast with CLOW [75] and DFR [74]. Although our results on the MT Defect dataset are slightly inferior to ViTALnet, our method still shows strong competitiveness and stability in terms of the average detection results of the KolektorSDD2 and MT datasets.

4) CONCLUSION

We agree that there is potential for optimization in the method used to convert pixel-relative results into image-

TABLE 1. Image-level / pixel-level AUROC(%) on MVTech AD dataset (Multi-class).

Category	PaDim[29]	MKD[72]	DRAEM[26]	JNLD[73]	UniAD[36]	PMAD[38]	Ours
Bottle	97.9 / 96.1	98.7 / 91.8	94.6 / 87.4	99.1 / 94.8	**100.0** / 96.2	**100.0 / 97.8**	**100.0**/ 96.9
Cable	70.9 / 81.0	78.2 / 89.3	61.8 / 70.4	90.6 / 76.4	94.2 / 95.3	**97.5 / 96.3**	97.3 / 93.1
Capsule	73.4 / **96.9**	68.3 / 88.3	70.2 / 49.2	74.4 / 57.0	89.9 / 92.5	91.2 / 96.2	**97.8** / 92.9
Carpet	93.8 / 97.6	69.8 / 95.5	95.9 / 95.2	77.1 / 93.7	99.5 / 97.4	**99.9 / 98.8**	98.4 / 98.2
Grid	73.9 / 71.0	83.8 / 82.3	98.1 / 99.0	98.6 / 96.9	98.4 / 94.8	98.2 / 96.2	**100.0 / 99.3**
Hazelnut	85.5 / 96.3	97.1 / 91.2	95.1 / 96.0	90.8 / 85.9	99.8 / 95.9	**100.0** / 98.0	99.2 / **99.4**
Leather	99.9 / 84.8	93.6 / 96.7	99.9 / 98.6	97.0 / 87.0	**100.0** / 97.3	**100.0** / 99.0	**100.0 / 99.5**
Metal nut	88.0 / 84.8	64.9 / 64.2	88.9 / 72.6	93.3 / **97.4**	99.5 / 92.5	**100.0** / 88.8	**100.0** / 97.4
Pill	68.8 / 87.7	79.7 / 69.7	69.0 / 90.0	82.7 / 91.2	93.7 / 97.9	96.5 / 95.2	**99.3 / 98.2**
Screw	56.9 / 94.1	75.6 / 92.1	93.3 / 89.3	81.8 / 87.0	87.5 / **96.8**	80.7 / 95.4	**95.6** / 93.4
Tile	93.3 / 80.5	89.5 / 85.3	98.3 / 98.1	99.2 / 94.7	99.3 / 98.7	**100.0** / 95.6	99.2 / **99.3**
Toothbrush	95.3 / 95.6	75.3 / 88.9	82.8 / 94.4	**100.0** / 98.6	94.2 / 96.5	89.4 / 98.0	99.7 / **99.4**
Transistor	86.6 / 92.3	73.4 / 71.7	83.9 / 73.1	90.3 / 83.6	**98.8 / 95.8**	96.3 / 94.0	95.6 / 91.7
Wood	72.1 / 89.1	93.4 / 80.5	99.8 / 96.2	91.9 / 88.7	98.6 / 91.8	**100.0** / 90.8	97.5 / **97.9**
Zipper	79.7 / 94.8	87.4 / 86.1	99.1 / 96.9	99.8 / 95.3	96.8 / 92.4	96.7 / 94.2	**100.0 / 99.2**
Average	82.4 / 89.5	81.9 / 84.9	88.7 / 87.1	91.1 / 88.5	96.7 / 95.5	96.4 / 95.6	**98.6 / 97.1**

TABLE 2. Image-level / pixel-level AUROC(%) on MVTech AD dataset (Single-class).

Method	PaDim[29]	MKD[72]	DRAEM[26]	JNLD[73]	UniAD[36]	PMAD[38]	Ours
Image-level AUROC(%)	95.2	89.1	98.1	97.4	96.6	97.4	**98.4**
Pixel-level AUROC(%)	97.4	90.7	97.3	**97.9**	96.5	97.8	97.2

TABLE 3. Image-level / pixel-level AUROC(%) on VisA dataset.

Category	PaDim[29]	JNLD[73]	UniAD[36]	OmniAL[37]	Ours
PCB1	92.7 / 97.7	71.3 / 98.6	82.0 / 96.4	**96.6 / 98.7**	91.9 / 91.6
PCB2	87.9 / 97.2	89.7 / 92.5	96.3 / 91.9	**99.4** / 83.2	85.5 / **97.6**
PCB3	85.4 / 96.7	73.1 / 93.8	**96.9** / 95.3	**96.9** / 98.4	94.9 / **99.5**
PCB4	**99.1** / 89.2	91.3 / 95.8	94.8 / 96.1	97.4 / **98.5**	98.8 / 98.2
Macaroni1	85.7 / 98.8	70.3 / 95.8	94.3 / 98.8	96.9 / **98.9**	**98.5** / 97.2
Macaroni2	70.8 / 96.0	71.3 / 94.1	86.5 / 92.9	**89.9 / 99.1**	85.0 / 96.6
Capsules	68.1 / 86.3	77.3 / 93.7	89.1 / **98.9**	87.9 / 98.6	**90.1** / 93.8
Candles	**89.1 / 97.3**	82.3 / 87.0	**89.1** / 94.8	85.1 / 90.5	88.3 / 94.4
Cashew	90.5 / 86.1	94.2 / 94.7	96.0 / 96.3	**97.1** / 98.9	95.7 / **99.1**
Chewing gum	**99.3** / 96.9	93.4 / 97.5	98.5 / **99.4**	94.9 / 98.7	96.1 / 99.0
Fryum	89.8 / 88.0	**100.0** / 97.5	93.2 / 95.8	97.0 / 89.3	96.9 / **99.1**
Pipe fryum	95.6 / 95.4	94.1 / 81.8	**96.0** / 97.0	91.4 / 99.1	94.9 / **99.7**
Average	87.8 / 93.8	84.0 / 93.6	92.7 / 96.1	**94.2** / 96.0	93.0 / **97.2**

TABLE 4. Ablation results of Image-level / pixel-level AUROC(%) on MVTech AD dataset.

Pipeline	Stage 1	75% masked in stage 1	PLPM	TBM	Results
Reconstruction-based					93.2 / 87.9
	✓				96.5 / 92.6
	✓	✓			95.3 / 91.8
	✓	✓	✓		96.8 / 95.8
End-to-End					94.7 / 89.3
	✓	✓	✓		96.1 / 92.2
	✓	✓		✓	95.2 / 93.7
	✓	✓	✓	✓	**98.6 / 97.1**

TABLE 5. Pixel-level AUROC(%) on KolektorSDD2 and MT datasets.

Method	KolektorSDD2	MT Defect	Average
DFR[74]	93.5	94.7	94.1
CLOW[75]	95.8	90.2	93.0
ViTALnet[66]	96.2	**99.0**	**97.6**
Ours	**98.7**	95.8	97.3

TABLE 6. Training / testing speed (images/sec) and learnable parameters (M) of the models.

Task	Methods	Training	Testing	Param
Single-class	PaDim[29]	32.5	39.5	68.8
	MKD[72]	35.3	42.9	24.9
	DRAEM[26]	23.5	41.5	97.5
Multi-class	UniAD[36]	38.6	64.7	27
	PMAD[38]	31.9	31.1	92.0
	Ours	25.2	37.7	84.8

level scores. The VisA dataset generally consists of smaller abnormal regions than the MVTec AD dataset. Consequently, lightweight networks may not achieve fine-grained results. Additionally, the uncertainty regions (gray or blurred parts of

the output maps) tend to be larger, indicating that a smaller network could help mitigate the influence of misleading information and provide a more comprehensive assessment of anomalies.

In general, to further validate the practicality of MALA, we use the most challenging public datasets, MVTec Anomaly Detection (MVTec AD) [22], Visual Anomaly Dataset(VisA) [24], KolektorSDD2 [43], and MT [44] to train. In brief, MALA achieves SOTA results in the MVTech AD and KolektorSDD2, achieves SOTA in the pixel-level of the ViSA dataset, and ranks second only to OmniAL in terms of image level, ranks second in MT Defect.

E. ABLATION STUDIES

In this section, we demonstrate the necessity and effectiveness of our training paradigm and enhanced modules. First, we attempt different training pipelines between Reconstruction-based and End-to-End. Next, we enhance the encoder's learning ability by masking images. Then, we utilize the Pseudo Label Prediction Module to improve the overall learning ability of the stage 1 model for defect details by processing the true and false labels. Finally, we fed the tokens into TBM to generate feature maps of different scales. All the experiments are based on the MVTec AD dataset.

As visualized in Table 4, the results of our ablation experiment clearly demonstrate that both the modules and retraining setups have a significant impact on the final performance of our anomaly localization and detection goal. **Training Pipeline**. The end-to-end training strategy can reduce manual pre-processing and subsequent processing and make the model go from original input to final output as much as possible, giving the model more space to automatically adjust according to the data and increasing the model's overall fitness. The first and fifth rows in the chart are different results obtained by different training pipelines. It can be seen that without adding other modules, end-to-end can get better results compared to the Reconstruction-based pipeline increased by 1.5%/1.4%. The use of **Stage 1** can better help the encoder capture relevant details. The difference between the first and second rows in the table is whether stage 1 is used during the training process. It can be seen that the accuracy of the model has been significantly improved after adding stage 1. In order to better allow the encoder to learn defect details, we added a **Pseudo Label Prediction Module**, which consists of convolutional layers and is trained to distinguish label types. The experiments in the third and fourth rows and the last two rows prove the effectiveness of PLPM. It can be seen that PLPM has indeed optimized the model forward, and the accuracy has been significantly improved. Finally, the first and third last rows in the table verify the role of **TBM**. The tokens are fed into TBM to generate feature maps of different scales, and the final result is a 2.5%/4.9% improvement.

F. COMPUTATION COST ANALYSIS

We analyze the cost of different models. All the values are measured with a single GPU (GeForce RTX 2080 Ti) on the MVTech AD dataset. The results are listed in Table 6. All the input images are resized into 256×256 and used to test the ability to process images per second and trainable parameters.

Although the multi-class models seem to contain more trainable parameters, single-class models indeed require more storage. For instance, when it comes to 15 classes, even the MKD [72] requires over 320M parameters, which are obviously much more than each of the multi-class models, leaving alone larger single-class models [26], [29]. Due to the larger pipelines, multi-class models may train and infer more slowly. Although UniAD [36] seems to perform perfectly on both time and space, the properly extracted feature tokens of each class are so challenging that transferability is limited. Compared with the SOTA PMAD [38], our model is competitive in both time and space.

V. DISCUSSION AND CONCLUSION

A. DISCUSSION

Despite our excellent results in many datasets and industrial scenarios, deployment of redundant parameters will inevitably cause the inference speed of the model to decrease, and the deployment will take more time compared to other models due to the need for the two-stage feature. However, the large number of parameters makes our model play a dual role by augmenting perceptual capabilities for larger models and standing independently as a trainable stand-alone model. Its unique strength lies in its ability to effectively detect diverse defects without relying on pre-existing parameters, showcasing adaptability to unique scenarios and unknown anomalies. In the realm of industrial defect detection, taking manufacturing and electronics as examples, our model, when deployed in cameras, proves adept at identifying surface defects, wear and tear, and soldered connections on circuit boards. This not only elevates product reliability but also enables timely adjustments for improved quality control. Beyond these advantages, the broader impact of our model is noteworthy. It has the potential to significantly reduce manpower requirements in industrial settings. This reduction translates to substantial cost savings, while concurrently enhancing safety measures for a more robust operational environment.

B. CONCLUSION

This paper presents a novel MAE-based self-supervised network called MALA, which leverages multi-stage pretext tasks of inpainting to accomplish final end-to-end anomaly detection and localization. Our pipeline achieves SOTA results in MVTech AD and VisA classes of image- and pixel-level AUROC. Besides, MALA also achieves considerable segmentation performance on the KolektorSDD2 and MT datasets. This progress indicates the innovation and robustness of MALA, even if the input datasets are made

up of various types and styles. However, a few indexes still have the distance to keep up with the apex indexes. Besides, the predicted anomaly maps may contain uncertain regions leading to misreporting. Due to the deep perception characteristic of ViT, the training outcome is satisfying, while the training process is time-consuming.

In the near future, we aim to combine multi-modal input with the training process to assess defective regions more precisely and study the more challenging multi-dataset setup. Besides, the distillation of parameters is also a promising research orientation to speed up the training and validation process while retaining excellent experimental results as much as possible.

REFERENCES

[1] X. Yao, R. Li, J. Zhang, J. Sun, and C. Zhang, "Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24490–24499.

[2] T. Xiang, Y. Zhang, Y. Lu, A. L. Yuille, C. Zhang, W. Cai, and Z. Zhou, "SQUID: Deep feature in-painting for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23890–23901.

[3] T. Liang, H. Bao, W. Pan, and F. Pan, "Traffic sign detection via improved sparse R-CNN for autonomous vehicles," *J. Adv. Transp.*, vol. 2022, pp. 1–16, Mar. 2022.

[4] T. Liang, H. Bao, W. Pan, X. Fan, and H. Li, "DetectFormer: Category-assisted transformer for traffic scene object detection," *Sensors*, vol. 22, no. 13, p. 4833, Jun. 2022.

[5] T. Wang, Z. Miao, Y. Chen, Y. Zhou, G. Shan, and H. Snoussi, "AED-Net: An abnormal event detection network," *Engineering*, vol. 5, no. 5, pp. 930–939, Oct. 2019.

[6] G. Yu, S. Wang, Z. Cai, X. Liu, C. Xu, and C. Wu, "Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13987–13998.

[7] K. V. Thakare, Y. Raghuwanshi, D. P. Dogra, H. Choi, and I.-J. Kim, "DyAnNet: A scene dynamicity guided self-trained video anomaly detection network," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5541–5550.

[8] A. Aich, K.-C. Peng, and A. K. Roy-Chowdhury, "Cross-domain video anomaly detection without target domain adaptation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2579–2591.

[9] S. Liu, S. Huang, X. Xu, J. Lloret, and K. Muhammad, "Efficient visual tracking based on fuzzy inference for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 20, 2023, doi: 10.1109/TITS.2022.3232242.

[10] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "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.

[11] D. Li, Y. Li, J. Li, and G. Lu, "PPR-Net: Patch-based multi-scale pyramid registration network for defect detection of printed labels," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 4061–4076.

[12] C. Lv, Z. Zhang, F. Shen, and F. Zhang, "Unsupervised automatic defect inspection based on image matching and local one-class classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4434–4443.

[13] H. Zhang, Z. Wu, Z. Wang, Z. Chen, and Y.-G. Jiang, "Prototypical residual networks for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16281–16291.

[14] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1–7.

[15] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

[16] B. Zuo and F. Wang, "Surface cutting defect detection of magnet using Fourier image reconstruction," *Comput. Eng. Appl.*, vol. 52, no. 3, pp. 256–260, 2016.

[17] Y. Xia, C. Luo, Y. Zhou, and L. Jia, "A hybrid method of frequency and spatial domain techniques for TFT-LCD circuits defect detection," *IEEE Trans. Semicond. Manuf.*, vol. 36, no. 1, pp. 45–55, Feb. 2023.

[18] Z. Gao, X. Zhao, M. Cao, Z. Li, K. Liu, and B. M. Chen, "Synergizing low rank representation and deep learning for automatic pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 10676–10690, Oct. 2023.

[19] C. Zhang, Y. Wang, and W. Tan, "MTHM: Self-supervised multi-task anomaly detection with hard example mining," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 3518613.

[20] X. Tao, S. Yan, X. Gong, and C. Adak, "Learning multi-resolution features for unsupervised anomaly localization on industrial textured surfaces," *IEEE Trans. Artif. Intell.*, early access, Dec. 6, 2022, doi: 10.1109/TAI.2022.3227142.

[21] D. Kim, C. Park, S. Cho, and S. Lee, "FAPM: Fast adaptive patch memory for real-time industrial anomaly detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[22] 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.

[23] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "VT-ADL: A vision transformer network for image anomaly detection and localization," in *Proc. IEEE 30th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2021, pp. 1–6.

[24] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 392–408.

[25] W.-H. Chu and K. M. Kitani, "Neural batch sampling with reinforcement learning for semi-supervised anomaly detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 751–766.

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

[27] 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.

[28] X. Tao, D. Zhang, W. Ma, Z. Hou, Z. Lu, and C. Adak, "Unsupervised anomaly detection for surface defects with dual-Siamese network," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7707–7717, Nov. 2022.

[29] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2021, pp. 475–489.

[30] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14318–14328.

[31] A. Yan, P. Rupnowski, N. Guba, and A. Nag, "Towards deep computer vision for in-line defect detection in polymer electrolyte membrane fuel cell materials," *Int. J. Hydrogen Energy*, vol. 48, no. 50, pp. 18978–18995, Jun. 2023.

[32] Q. Wan, L. Gao, X. Li, and L. Wen, "Industrial image anomaly localization based on Gaussian clustering of pretrained feature," *IEEE Trans. Ind. Electron.*, vol. 69, no. 6, pp. 6182–6192, Jun. 2022.

[33] 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.

[34] O. Rippel, P. Mertens, and D. Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6726–6733.

[35] L.-L. Chiu and S.-H. Lai, "Self-supervised normalizing flows for image anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2926–2935.

[36] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A unified model for multi-class anomaly detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 4571–4584.

[37] Y. Zhao, "OmniAL: A unified CNN framework for unsupervised anomaly localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3924–3933.

[38] X. Yao, C. Zhang, R. Li, J. Sun, and Z. Liu, "One-for-all: Proposal masked cross-class anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 4792–4800.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

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

[41] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[42] W. Jin, F. Guo, and L. Zhu, "ISSTAD: Incremental self-supervised learning based on transformer for anomaly detection and localization," 2023, *arXiv:2303.17354*.

[43] 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.

[44] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *Vis. Comput.*, vol. 36, no. 1, pp. 85–96, Jan. 2020.

[45] S. Sheynin, S. Benaim, and L. Wolf, "A hierarchical transformation-discriminating generative model for few shot anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8495–8504.

[46] 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.

[47] N. Belton, M. T. Hagos, A. Lawlor, and K. M. Curran, "FewSOME: One-class few shot anomaly detection with Siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2977–2986.

[48] T. Aota, L. T. T. Tong, and T. Okatani, "Zero-shot versus many-shot: Unsupervised texture anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5564–5572.

[49] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[50] H. Yao, W. Yu, W. Luo, Z. Qiang, D. Luo, and X. Zhang, "Learning global–local correspondence with semantic bottleneck for logical anomaly detection," 2023, *arXiv:2303.05768*.

[51] P. Nooralishahi, G. Ramos, S. Pozzer, C. Ibarra-Castanedo, F. Lopez, and X. P. V. Maldague, "Texture analysis to enhance drone-based multi-modal inspection of structures," *Drones*, vol. 6, no. 12, p. 407, Dec. 2022.

[52] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, and C. Wang, "Multimodal industrial anomaly detection via hybrid fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8032–8041.

[53] Y. Cao, X. Xu, C. Sun, Y. Cheng, L. Gao, and W. Shen, "2nd place winning solution for the CVPR2023 visual anomaly and novelty detection challenge: Multimodal prompting for data-centric anomaly detection," 2023, *arXiv:2306.09067*.

[54] J. Wang, G. Xu, F. Yan, J. Wang, and Z. Wang, "Defect transformer: An efficient hybrid transformer architecture for surface defect detection," *Measurement*, vol. 211, Apr. 2023, Art. no. 112614.

[55] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi, "Just noticeable difference estimation for images with free-energy principle," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1705–1710, Nov. 2013.

[56] M. Yang, P. Wu, and H. Feng, "MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105835.

[57] W. Luo, H. Yao, and W. Yu, "Normal reference attention and defective feature perception network for surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[58] C. Ding, G. Pang, and C. Shen, "Catching both gray and black swans: Open-set supervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7388–7398.

[59] J. Long, Y. Yang, L. Hua, and Y. Ou, "Self-supervised augmented patches segmentation for anomaly detection," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1926–1941.

[60] V. Zavrtanik, M. Kristan, and D. Skočaj, "DSR—A dual subspace re-projection network for surface anomaly detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 539–554.

[61] J. Pirnay and K. Chai, "Inpainting transformer for anomaly detection," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2022, pp. 394–406.

[62] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14668–14678.

[63] C. Zhang, W. Dai, V. Isoni, and A. Sourin, "Automated anomaly detection for surface defects by dual generative networks with limited training data," *IEEE Trans. Ind. Informat.*, early access, Mar. 31, 2023, doi: 10.1109/TII.2023.3263517.

[64] N.-C. Ristea, N. Madan, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised predictive convolutional attentive block for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13576–13586.

[65] H. Liu, X. Miao, C. Mertz, C. Xu, and H. Kong, "CrackFormer: Transformer network for fine-grained crack detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3783–3792.

[66] X. Tao, C. Adak, P.-J. Chun, S. Yan, and H. Liu, "ViTALnet: Anomaly on industrial textured surfaces with hybrid transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[67] X. Xie, Y. Huang, W. Ning, D. Wu, Z. Li, and H. Yang, "RDAD: A reconstructive and discriminative anomaly detection model based on transformer," *Int. J. Intell. Syst.*, vol. 37, no. 11, pp. 8928–8946, Nov. 2022.

[68] X. Y. Lee, L. Vidyaratne, M. Alam, A. Farahat, D. Ghosh, T. G. Diaz, and C. Gupta, "XDNet: A few-shot meta-learning approach for cross-domain visual inspection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4374–4383.

[69] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19606–19616.

[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[71] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[72] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14902–14912.

[73] Y. Zhao, "Just noticeable learning for unsupervised anomaly localization and detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

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

[75] 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.

YIBO CHEN received the B.Eng. degree in automation from Zhejiang University, Hangzhou, China, where he is currently pursuing the master's degree with the College of Control Science and Engineering. He is with the College of Control Science and Engineering. His research interests include industrial anomaly detection, pattern recognition, and knowledge distillation.

HAOLONG PENG received the B.Eng. degree in 2022. He is currently pursuing the master's degree with the Polytechnic Institute, Zhejiang University, Hangzhou, China. During his postgraduate study, he participated in many on-campus scientific research activities and achieved excellent results. His research interests include 3D segmentation, masked autoencoders, multimodal fusion, and knowledge distillation.

LE HUANG is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include anomaly detection, especially surface defect detection of industrial products, face recognition/anti-spoofing, and 3D reconstruction.

JIANMING ZHANG received the Ph.D. degree. He is currently an Associate Professor with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China. He has long been engaged in artificial intelligence, intelligent control and process optimization, process monitoring, information integration, data mining, and robotics.

WEI JIANG received the Ph.D. degree in pattern recognition from the Tokyo Institute of Technology, Tokyo, Japan. He is currently an Associate Professor with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include multimedia analysis, large-scale pattern recognition, computer vision, deep learning, and control systems.