Discriminative Feature Learning Framework With Gradient Preference for Anomaly Detection

Muhao Xu, Xueying Zhou, Xizhan Gao, WeiKai He, and Sijie Niu, Member, IEEE

Abstract—Unsupervised anomaly detection methods can detect product defects in industrial images by leveraging only normal samples during model training. Currently, the representation-based method, as a popular unsupervised anomaly detection method, has achieved impressive performance. Extracting valuable feature vectors that can remarkably improve the performance of anomaly detection is essential in unsupervised representation learning. To this end, we propose a novel discriminative feature learning framework with gradient preference for anomaly detection. Specifically, a gradient preference-based selector is designed to construct a representative feature repository, which alleviates the interference of the redundant feature vectors. To further make the feature vectors of normal samples adapt to the target distribution, a discriminative feature learning framework with center constraint is presented. By introducing our proposed framework, the distribution of normal samples becomes more compact. In the inference stage, the anomaly score calculated by the distance between feature vectors of test data and the normal cases is used to detect anomalies. Moreover, our method can be easily extended to anomaly localization. Extensive experiments on three popular industrial anomaly detection datasets demonstrate our proposed framework can achieve competitive results in both anomaly detection and localization. Additionally, the results on two medical anomaly detection datasets validate the generalization performance of our proposed method.

Index Terms—Anomaly detection, discriminative feature learning, feature vector selection, gradient preference, unsupervised learning.

I. INTRODUCTION

ANOMALY detection [1], [2], [3] refers to identifying whether unseen data have expected behavior or not, which has been extensively studied in various fields ranging from industrial image-based defects detection [4], [5], [6], [7], [8], [9], [10] and medical image-based disease diagnosis [11], [12], [13], [14]. Some defects in industry, such as scratches or dents on the surface of products, the absence of certain product parts and the distorted product structure, are usually considered anomalies. Automatic detection of these anomalies is an essential part of industrial intelligence, which is conducive to ensuring product quality and user experience. Generally, challenges arising in anomaly detection are as follows: 1) As the types of anomalies are complicated and various, it is difficult to collect all of them. 2) Abnormal samples in the vast majority of datasets are very rare. As a consequence, unsupervison-based methods are becoming the mainstream techniques in anomaly detection [15], [16].

Unsupervised anomaly detection methods are desirable to train models solely on anomaly-free images. Existing works predominantly focus on reconstruction-based methods, such as auto-encoder [17], [18], [19] or generative adversarial networks (GAN) [13], [14], [20]. The principle of these methods is to detect anomalies with per-pixel reconstruction error or by evaluating the density obtained from the model’s probability distribution. Although features learned by the reconstruction-based methods are beneficial to pixel reconstruction, it is difficult for these methods to capture features that are conducive to anomaly detection.

Recently, representation-based methods [9], [21], [22], [23], [24], [25], [26], [27], [28], [29] have achieved remarkable success in anomaly detection. These methods extract the deep representations (i.e., feature vectors or patches) of the samples by the pretrained networks, and detect anomalies through representation or distribution distance. A simple strategy proposed by previous methods [21], [30] is to memorize the representations of normal samples in the training phase, and detect anomalies through the representation difference between test samples and memorized normal ones. However, due to little consideration of anomaly detection tasks in the process of deep representation extraction, it is difficult to separate abnormal samples from normal ones, such as anomaly cases 1 and 2 in Fig. 1(b). Therefore, feature adaptation is necessary for better anomaly detection performance. To achieve that, recent approaches retrain the pretrained networks on the normal samples to adapt the target distribution. For instance, Panda [22] retrain the pretrained network with one-class constraint and detects anomalies by feature distance between normal data in the training dataset and test data.

Although the previous works have achieved considerable success, some challenging problems still need to be addressed. 1) Most representation-based methods employ the deep representations of normal samples without any selection to model the distribution of normal cases. Nevertheless, there are a lot of redundancies in these deep representations, resulting in limited anomaly detection performance.
1) We propose a gradient preference-based feature vector selection strategy to alleviate the interference of redundant feature vectors, ensuring the representativeness of the modeled distribution of normal samples.

2) Different from previous feature adaptation methods [22], [23], we construct a discriminative feature learning network to make the distribution of normal samples so compact that the normal and abnormal samples are more separable. Furthermore, an early stop strategy is given to avoid the “mode collapse” problem.

3) Extensive experiments are conducted on five popular industrial and medical anomaly detection datasets. The results demonstrate that our proposed method can achieve competitive performance in both anomaly detection and localization.

II. RELATED WORK

Classical anomaly detection methods focus on learning a boundary to separate the abnormal samples from the normal ones. Schölkopf et al. [31] proposed to construct a one-class support vector machine (OCSVM) to learn a discriminative hyperplane that can distinguish normal and abnormal samples. Li et al. [32] propose to use the Gaussian mixture model (GMM) to model the distribution of normal samples so that the anomalies can be detected due to being far away from the learned distribution. With the development of deep learning, Ruff et al. [33] propose a deep one-class support vector data description (Deep-SVDD) to learn a hypersphere by minimizing the distance between feature vectors of normal images extracted by a neural network. Though many efforts, it is easy for the model to overfit the training data because of a lack of sufficient generalization constraints. Later, several methods, including the patch-level SVDD [34] and interpolated Gaussian descriptor (IGD) method [12], are proposed to overcome the above challenge.

Reconstruction-based methods have been widely explored in anomaly detection, which detects anomalies with the assumption that a model trained only with normal images cannot reconstruct abnormal images very well. These methods usually rely on encoder–decoder scheme, and use the reconstruction error as the metric to detect anomalies. Akçay et al. [35] propose an encoder-decoder convolutional neural network, and detect anomalies by combining feature and image reconstruction errors. However, the assumption that abnormal images are difficult to be reconstructed well does not always hold [14], [36]. To overcome that, many works [10] impose constraints to obtain a more compact feature distribution of normal images, which is beneficial to improve the sensitivity of models to anomalies. Schlegl et al. [37] train a GAN to learn the distribution of normal image patches. In the test phase, the image patches of test data are constructed into normal ones so that the anomalies can be detected due to large reconstruction errors. Similarly, You et al. [38] train a variational auto-encoder (VAE) to model the distribution of normal images by introducing Kullback-Leibler divergence in feature space, which can reconstruct abnormal images into images with normal anatomy and detect anomalies by reconstruction error. Additionally, memory network [36], [39], [40] is one of the common reconstruction models for anomaly detection. In the training phase, it stores the features of normal samples in memory modules by minimizing their reconstruction error and reconstructs test samples using the stored features in the test phase. Therefore, the abnormal samples are captured due to large reconstruction errors between test data and their reconstruction ones.
**Representation-based methods** achieve excellent performance in both anomaly detection and localization. These methods extract representations (i.e., feature vectors or patches) of images from the pretrained network, and detect anomalies by the distance between the representation of test images and normal cases. Teacher–student network is a classical framework for representation-based anomaly detection methods. These studies [9], [24], [25], [26], [29] distill the knowledge of normal samples from the pretrained teacher network into a student network, which makes the network sensitive to the abnormal samples that are obviously different in the teacher and the student network. In addition, some works [21], [22], [30] memorize the representations of normal samples and detect anomalies by the difference between the representations of test samples and the memorized ones. PaDiM [21]memorizes the multivariate Gaussian distributions of patch embeddings of the normal class extracted by the pretrained network, and detects anomalies by the distance between patch embeddings of test data and the memorized distribution. Semantic pyramid anomaly detection (SPADE) [30] stores the feature vectors of normal samples extracted by the pretrained network, and uses $K$-nearest neighbor between feature vectors of test data and the stored ones to detect anomalies. To further improve the anomaly detection performance of SPADE, an improved version, Panda [22], proposes to employ one-class constraint to retrain the pretrained network to adapt to the target distribution. Additionally, FYD [23] retrains the pretrained network with non-contrastive learning framework, which makes the representations of normal samples closer. In the inference phase, the anomalies can be captured due to the deviation from the learned distribution. Furthermore, the flow model, which can estimate the density of given samples, has also been extensively studied in anomaly detection. DifferNet [27] leverages a multiscale feature extractor based on a pretrained network to obtain meaningful features of normal samples, and estimate their density using normalizing flows. CS-Flow [28] proposes a novel fully convolutional cross-scale normalizing flow that jointly processes multiple feature maps of different scales. In the inference phase, these methods detect anomalies by estimating the likelihoods of test samples.

### III. Method

To make the normal and abnormal samples more distinguishable, it is necessary to consider a discriminative feature learning network for bringing the features obtained by the pretrained network close to better anomaly detection results. In the training stage, we first obtain multilevel aggregation features of normal samples by combing different level features extracted by the pretrained network. Then, a gradient preference-based feature vector selection strategy is proposed to reduce the interference of redundant feature vectors. Meanwhile, we propose a discriminative feature learning framework with center constraint to make the selected feature vectors adapt to the target distribution. In the test phase, the distance between feature vectors of test data and those in feature repository is used to calculate the image- and pixel-level anomaly score, which are defined as the metric for anomaly detection and localization. Moreover, the architecture of our proposed method is illustrated in Fig. 2.

#### A. Feature Extraction With Pretrained Network

The pretrained networks are able to capture different level features in a supervised manner, which can be transferred to anomaly detection tasks. In our work, we first obtain a pretrained network $\phi$, such as ResNet and EfficientNet, trained on an additional dataset (e.g., ImageNet). Then, our training images $x_i \in \mathbb{R}^{c \times h \times w}$ (channel $c$, height $h$ and width $w$) are fed into the pretrained network $\phi$, and the different level features $f_{i,l}$ are defined as

$$f_{i,l} = \phi(x_i)$$

where $l$ denotes different levels, and $x_i$ denotes $i$th training sample.

In general, the low-level features represent the edge and texture of images, while the high-level features represent the semantic information of images. Hence, we combine features of different levels to get the final multilevel aggregation feature $F_i \in \mathbb{R}^{C \times H \times W}$ of each image $x_i$, which can be written as

$$F_i = \text{AP}(f_{i,1}) \oplus \text{AP}(f_{i,2}) \oplus \cdots \oplus \text{AP}(f_{i,l})$$

where $\oplus$ represents concatenation in channel dimension, and $\text{AP}(\cdot)$ represents adaptive pooling that can resize $f_{i,j}$ into the resolution of maximum one. Finally, we obtain the feature vectors of pixel location $(j, k)$ in $F_i$, defined as $F_i(j, k) \in \mathbb{R}^C$, where $j \in [0, W)$ and $k \in [0, H)$.

#### B. Gradient Preference Based Feature Vector Selection

The feature vectors obtained by the previous step can be directly employed to anomaly detection. However, the
performance is limited as there are a lot of redundancies in these feature vectors. Also, as the size of the training dataset increases, the inference time and the required storage also increase. To overcome the above problems, we propose a novel selection strategy based on gradient preference to select feature vectors that are beneficial to anomaly detection.

Concretely, inspired by image sharpening, we calculate the gradient on the multilevel aggregation features to generate the gradient score map $S_i \in \mathbb{R}^{H \times W}$, defined as

$$S_i = \Xi(F_i \odot G_E)$$

where $S_i(j, k)$ indicates the gradient value of feature vector $F_i(j, k)$, and $\odot$ is the convolution operation, and $\Xi(.)$ is the normalization function, and $G_E \in \mathbb{R}^{C \times k \times k}$ is the convolution kernel obtained by expanding gradient operator $G \in \mathbb{R}^{k \times k}$. Different from the first-order differential operator, the Laplacian operator poses a strong ability to perceive texture changes [41], which is beneficial to detecting anomalies. Thus, Laplacian operator is employed to extract feature gradient, defined as

$$G = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}.$$  \hspace{1cm} (4)

The most straightforward selection strategy is to sort score map from large to small and select feature vectors with high gradient values, whereas in fact that feature vectors without high gradient increases, the inference time and the required storage also increase. To overcome the above problems, we propose a novel selection strategy based on gradient preference to select feature vectors with high gradient score map from large to small and select feature vectors with high gradient values, whereas in fact that feature vectors without high gradient values are also significant for anomaly detection. Therefore, it is hoped that feature vectors with high gradient values are selected with high probability, and those with low gradient values can also be considered. Thus, selecting feature vector $F_i(j, k)$ to be stored is actually a Bernoulli event that happens with probability $P = S_i(j, k)$. Concretely, we generate a random tensor $R_i \in \mathbb{T}^{H \times W}$, where $\mathbb{T} \sim U(0, 1)$. Then, the mask $B_i \in \mathbb{R}^{H \times W}$ based on gradient preference is defined as the following expression:

$$B_i(j, k) = \begin{cases} 1, & S_i(j, k) \geq R_i(j, k) \\ 0, & S_i(j, k) < R_i(j, k). \end{cases}$$

If $B_i(j, k) = 1$, the corresponding feature vectors $F_i(j, k)$ are selected to be stored in a feature repository $M \in \mathbb{R}^{M \times C}$, whose formula can be written as

$$M = \bigcup_{(j, k) \in \{(j, k) \mid B_i(j, k) = 1\}} F_i(j, k).$$

### C. Discriminative Feature Learning With Center Constraint

Although the above steps make the feature repository of normal samples more representative, there are still some problems to be solved. For one thing, the feature vectors we selected are redundant in dimension, which leads to inaccurate feature representation. For another thing, as shown in Fig. 1(b), the feature distribution of normal samples without any adaptation is so loose that the normal and abnormal samples are indistinguishable.

To alleviate the above problems, we utilize a simple multilayer perceptron (MLP) network $\varphi$ with central constraint loss to map the feature vectors of normal samples into a compact subspace with $C_f$ dimension, so that the normal and abnormal samples in test dataset are more discriminative, as reported in Fig. 1(c). Our proposed central constraint loss is as follows:

$$L_{\text{center}} = \sum_{i=1}^{M} \left| \frac{1}{M} \sum_{\varphi(\nu) \neq \varphi(\nu')} \right|_{1}$$

where $M$ denotes the number of feature vectors $\nu$ in feature repository $M$.

Under the condition of full training, the network $\varphi$ suffers “mode collapse” that all feature vectors are mapped to the same output and become inseparable. This phenomenon has been extensively observed in previous works [22], [23], [42] due to the lack of negative samples in training stage. To avoid that, we design an early stopping strategy. Specifically, we set a hyperparameter $\eta$, and stop the optimization of the network $\varphi$ when the given condition $C$ is met. The form of the condition $C$ is as follows:

$$L_{\text{center}} \leq \eta * L_{\text{start}}$$

where $L_{\text{start}}$ represents the loss values after training one epoch. Finally, we obtain the final feature repository $M_f \in \mathbb{R}^{M \times C_f}$ by

$$M_f = \bigcup_{\nu \in M} \varphi(\nu).$$

### D. Anomaly Detection in Inference Stage

In inference phase, we use anomaly score as the metric to detect and localize anomalies through the following procedures.

**Anomaly detection** refers to assessing whether a test image is abnormal or not. To this end, we feed a test image $x_t \in \mathbb{R}^{C \times H \times W}$ into the feature extraction module, which yields $W \times H$ feature vectors $F_t(j, k)$. Next, all feature vectors without any selection are fed into the trained MLP network $\varphi$, and the feature repository of the test image $M_t \in \mathbb{R}^{(W \times H) \times C_f}$ without selection is obtained by

$$M_t = \bigcup_{(j, k) \in \{(j, k) \mid j \in [0, W), k \in [0, H]\}} \varphi(F_t(j, k)).$$

Then, we calculate the image-level anomaly score $A_{\text{image}}$ of the test image by

$$A_{\text{image}} = \max_{v_t \in M_t} \left( \text{dist}(v_t, M_f) \right)$$

where $\text{dist}(v_t, M_f)$ is defined as minimum $L_1$ distance between $v_t$ and each feature vector in feature repository $M_f$. A high image-level anomaly score $A_{\text{image}}$ indicates that the image is likely to be abnormal. Since all the feature vectors stored in the feature repository $M_f$ are from normal images, the features vectors of normal images in the test dataset are very close to those in feature repository $M_f$. In contrast, the feature vectors of abnormal regions have obvious difference from those in the feature repository $M_f$, leading to high anomaly scores.

**Anomaly localization** refers to judging whether a pixel is in an abnormal region or not. Similarly, we calculate anomaly score $A_{\text{pixel}}$ for pixels to indicate the possibility that the pixels
Algorithm 1 Discriminative Feature Learning Framework With Gradient Preference

**Input**: train data $x_t$, test data $x_t$

**Output**: image anomaly score $\mathcal{A}_{image}$, pixel anomaly score $\mathcal{A}_{pixel}$

**Initialize**: hyper-parameters $\eta, \alpha, \gamma$. pretrained network $\phi$.

1. **Training**
2. $f_{t,j} \leftarrow \phi(x_t)$
3. Combine multilevel features using (2) to get $F_t$
4. Calculate the gradient score map $S_t$ with (3)
5. $R \sim$ uniform (0,1)
6. $M \leftarrow \{\}$
7. if $S_t(j,k) \geq R(t,j,k)$ then
8. $M \leftarrow M \cup F_t(j,k)$
9. end if
10. while $\text{Loss} < \eta * \text{Loss}_{start}$ do
11. Training discriminative feature learning networks $\phi$
12. end while
13. $M_f \leftarrow \{\}$
14. $M_f \leftarrow M_f \cup \varphi(v); v \in M$
15. **Testing**
16. $f_{t,j} \leftarrow \phi(x_t)$
17. Combine multilevel features using (2) to get $F_t$
18. $M_t \leftarrow \{\}$
19. $M_t \leftarrow (j,k) \in [(0,0),(H,W)] \varphi(F_t(j,k))$
20. Calculate anomaly score using (11) and (12)
21. return image anomaly score $\mathcal{A}_{image}$, pixel anomaly score $\mathcal{A}_{pixel}$

are in abnormal regions. The anomaly score $\mathcal{A}_{pixel}$ can be represented as

$$\mathcal{A}_{pixel} = \text{dist}(v_t, M_t), \quad v_t \in M_t.$$ (12)

Then, we get an attention map $A \in \mathbb{R}^{h \times w}$, which concentrates on abnormal regions in an image. Finally, we resize this attention map into the resolution of the original image by bilinear interpolation, and smooth it with a Gaussian kernel $\sigma = 4$. The pseudo-code for our proposed method is shown in Algorithm 1.

**IV. EXPERIMENTS**

To validate the effectiveness of our proposed method, extensive experiments are conducted on five publicly-available industrial and medical anomaly detection datasets.

**A. Dataset**

**MVTec AD dataset** [45] is one of the most challenging industrial anomaly detection datasets with multiobject and multidefect data, containing image- and pixel-level labels. It consists of ten objects and five texture categories, and each category contains 60–320 training samples and about 100 test samples including one to eight types of defects.

**Magnetic Tile Defects dataset** (MTD) [8] has a total of 1344 magnetic tile images with image- and pixel-level annotations. It contains 952 defect-free images and 392 images with five types of defects, which are blowhole, crack, fray, break, and uneven. Following [27], 762 defect-free images are selected as training data, and 190 defect-free images and 392 defect images are used for testing.

**BrainTech AD dataset** (BTAD) [46] contains three kinds of industrial products showcasing body and surface defects with both image- and pixel-level labels. Each kind is composed of a training dataset with 400, 399, and 1000 normal images and a test dataset with 21/49, 30/200, and 410/31 normal/abnormal images, respectively.

**Head computed tomography (CT) dataset** [24] consists of 100 normal head CT images and 100 hemorrhagic head CT images. We perform five-fold cross-validation in this dataset and use 80 normal images as the training data and the rest as the test data in each validation.

**Brain MRI dataset** [47] is a brain MRI tumor detection dataset containing four classes: glioma, meningioma, no tumor, and pituitary. The 1595 images with no tumor in training dataset is used for the anomaly detection model training, and all the test data consisting of 405, 300, 306, and 300 images with no tumor, glioma, meningioma, and pituitary are used for model testing.

**B. Experimental Settings**

We use ResNeXt-101 pretrained on ImageNet as our backbone network, and define the outputs of conv two-five blocks as one–four level features, respectively. In the following experiments, we use one–three level features to calculate $F_t$. Our MLP network $\varphi$ consists of two fully connected layers with output channel of 1024 and 512, respectively, and we use the Adam optimizer with a learning rate of $1 \times 10^{-4}$ to train it. In addition, the hyperparameter $\eta$ of our early stop strategy is set to 0.8. It should be noted that our method is end-to-end in both training and inference phases, and only the parameters of the MLP network $\varphi$ are updated in the training stage. Furthermore, all images of the above datasets are resized to $224 \times 224$ pixels. For anomaly detection and localization, we adopt image- and pixel level area under the receiver operating characteristic curve (AUROC) as the evaluation metric. All the results are the average score of 30 runs.

To evaluate our proposed method, some classical and SOTA methods are involved for comparison.

1. **GANomaly** [17] is a classical reconstruction-based method. It trains an encoder-decoder-encoder framework by both image and feature reconstruction error, and detects anomalies by feature reconstruction error.

2. **Patch SVDD** [34] and **IGD** [12] are one-class-classification (OCC) model based methods that learn a stable discriminative hypersphere and constrain normal samples within it.

3. **U-Student** [9], **MKD** [24], **STFPM** [29], and **reverse distillation (RD)** [25] transfer the knowledge of normal samples from the pretrained teacher network to a simple student network.

4. **FYD** [23] leverages the pretrained network as the encoder of noncontrastive learning framework and retraining the framework by minimizing the distance between normal samples.
TABLE I

| Method        | Category | Copper | Glass | Leather | Tile | Wood | Bottle | Cable | Cascade | Hazelnut | Metal Nut | Pill | Bow | Bracelet | Watch | Zipper | Avg |
|---------------|----------|--------|-------|---------|------|------|--------|-------|---------|----------|-----------|------|-----|----------|-------|--------|-----|
| SVDD [14]     | 92.82    | 84.91  | 92.60 | 74.40   | 95.60| 95.90| 96.60  | 99.06 | 96.60   | 99.06    | 99.06    | 99.06| 99.06| 99.06    | 99.06| 99.06  | 99.06|
| BDG [12]      | 92.86    | 87.57  | 95.60 | 95.60   | 95.60| 95.60| 95.60  | 95.60 | 95.60   | 95.60    | 95.60    | 95.60| 95.60| 95.60    | 95.60| 95.60  | 95.60|
| CutPaste [9]  | 93.30    | 93.30  | 95.60 | 95.60   | 95.60| 95.60| 95.60  | 95.60 | 95.60   | 95.60    | 95.60    | 95.60| 95.60| 95.60    | 95.60| 95.60  | 95.60|
| MVTec [24]    | 76.59    | 78.09  | 78.09 | 78.09   | 78.09| 78.09| 78.09  | 78.09 | 78.09   | 78.09    | 78.09    | 78.09| 78.09| 78.09    | 78.09| 78.09  | 78.09|
| STEPD [19]    | 78.09   | 78.09  | 78.09 | 78.09   | 78.09| 78.09| 78.09  | 78.09 | 78.09   | 78.09    | 78.09    | 78.09| 78.09| 78.09    | 78.09| 78.09  | 78.09|
| FYD [23]      | 98.89    | 100.00 | 99.60 | 99.60   | 99.60| 99.60| 99.60  | 99.60 | 99.60   | 99.60    | 99.60    | 99.60| 99.60| 99.60    | 99.60| 99.60  | 99.60|
| DRAEM [41]    | 97.09    | 97.09  | 97.09 | 97.09   | 97.09| 97.09| 97.09  | 97.09 | 97.09   | 97.09    | 97.09    | 97.09| 97.09| 97.09    | 97.09| 97.09  | 97.09|
| CutPaste [44] | 95.89    | 95.89  | 95.89 | 95.89   | 95.89| 95.89| 95.89  | 95.89 | 95.89   | 95.89    | 95.89    | 95.89| 95.89| 95.89    | 95.89| 95.89  | 95.89|
| DiffNet [27]  | 82.38    | 84.29  | 87.11 | 94.28   | 89.81| 89.81| 89.81  | 89.81 | 89.81   | 89.81    | 89.81    | 89.81| 89.81| 89.81    | 89.81| 89.81  | 89.81|
| CS-Flow [28]  | 95.89    | 95.89  | 95.89 | 95.89   | 95.89| 95.89| 95.89  | 95.89 | 95.89   | 95.89    | 95.89    | 95.89| 95.89| 95.89    | 95.89| 95.89  | 95.89|
| Panda [22]    | 89.58    | 89.58  | 89.58 | 89.58   | 89.58| 89.58| 89.58  | 89.58 | 89.58   | 89.58    | 89.58    | 89.58| 89.58| 89.58    | 89.58| 89.58  | 89.58|
| PaDiM [21]    | 98.89    | 98.89  | 98.89 | 98.89   | 98.89| 98.89| 98.89  | 98.89 | 98.89   | 98.89    | 98.89    | 98.89| 98.89| 98.89    | 98.89| 98.89  | 98.89|

Average: 99.06

C. Experimental Results

1) Results on MVTec AD Dataset: We compare our method with the classical and state-of-the-art (SOTA) methods on MVTec AD dataset, and the quantitative results are shown in Table I. The results of GANomaly are adopted from [10], and the others are adopted from their original papers directly. For anomaly detection, our method achieves the best performance in five categories and surpasses the SOTA methods by 0.4% in terms of the average image-level AUROC (99.1%). For anomaly localization, our proposed method achieves SOTA performance in five categories and competitive performance in the average pixel-level AUROC (97.9%), which is only 0.3% lower than FYD. Nevertheless, our method is 1.4% higher than FYD in anomaly detection. It can be observed that, compared with the reconstruction models (i.e., GANomaly) and OOC models (i.e., Patch SVDD and IGD), representation-based methods achieve better anomaly detection performance. Although DRAEM and CutPaste create abnormal samples to improve the model ability of identifying abnormal images, their performance improvement is limited due to unpredictable variations in anomalies. Furthermore, our method can achieve competitive performance in both anomaly detection and localization only by a simple gradient preference based discriminative feature learning framework. In addition, we visualize the performance of our method for anomaly localization in Fig. 3. It can be seen that our method can well localize defects in both object and texture.

2) Results on MTD and BTAD Datasets: To further evaluate our proposed method, we conduct widely experiments on two other industrial datasets, which are MTD and BTAD dataset, and the quantitative results are reported in Table II. It can be seen that the representation-based methods achieve better performance like results on the MVTec AD dataset. For the MTD dataset, our method achieves the best anomaly detection/localization performance in the image/pixel-level AUROC.
(99.1%/84.7%), which is 0.3%/1.5% higher than the SOTA method. Similarly, for BTAD dataset, our method achieves the SOTA performance in average image/pixel-level AUROC (96.2%/97.6%) for anomaly detection/localization, which is 0.8%/0.1% improvement than the SOTA method. Furthermore, the qualitative results are reported in Fig. 4, which demonstrates that the regions our proposed method focuses on are highly consistent with the ground truth (GT).

3) Results on Head CT and Brain MRI Dataset: To validate the generalization of our proposed method in medical datasets, extensive experiments are conducted on the Head CT and brain MRI dataset for anomaly detection, and the quantitative results are shown in Table III. As can be seen from Table III, some methods that perform well on industrial anomaly detection datasets cannot generalize well on medical anomaly detection datasets, such as DRAEM, CutPaste, and RD. In contrast, our method achieves SOTA performance on Head CT dataset in image-level AUROC (83.7%), which outplays all SOTA methods by a huge margin (1.8%). At the same time, our proposed method also achieves excellent performance on brain MRI dataset in image-level AUROC (99.9%). The excellent performance proves that our method can detect tumors in brain MRI images well. Although these two medical datasets do not provide manual annotation of lesion regions, we show the qualitative results of different types of diseases in Fig. 5 to further illustrate the superior performance of our proposed method on medical datasets. In total, all the above experimental results demonstrate that our proposed method...
TABLE IV
QUANTITATIVE RESULTS WITH DIFFERENT PRETRAINED NETWORK ON MVTec AD DATASET FOR ANOMALY DETECTION/LOCALIZATION, AS MEASURED ON AUROC [%]. THE BEST RESULTS ARE MARKED IN BOLD

| Dataset | ResNet18 [48] | Wide_ResNet50 [49] | ResNeXt-101 [50] | EffiicientNet_b4 [51] |
|---------|---------------|---------------------|------------------|---------------------|
| Bag     | 59.4/69.3     | 66.6/89.3           | 94.3/38.9        | 90.5/98.4           |
| Gold    | 56.6/97.7     | 96.5/97.5           | 99.5/98.5        | 96.7/96.5           |
| Leather | 99.9/68.9     | 100/98.1            | 100/98.0         | 100/98.3            |
| Tile    | 89.3/33.5     | 99.9/93.9           | 99.6/95.9        | 99.6/93.4           |
| Wood    | 89.3/33.5     | 88.9/97.2           | 99.5/95.0        | 88.5/93.4           |
| Rattle  | 100/87.9      | 100/90.0            | 100/98.7         | 100/98.1            |
| Cable   | 91.3/41.4     | 88.5/97.2           | 97.6/94.9        | 98.9/97.7           |
| Capsule | 95.9/48.6     | 88.1/98.3           | 97.3/96.9        | 95.0/98.7           |
| Harmer  | 100/98.5      | 100/98.0            | 100/98.8         | 99.9/98.1           |
| Metal   | 99.4/97.0     | 99.2/95.9           | 100/97.6         | 99.5/97.3           |
| Rock    | 95.3/98.2     | 95.7/98.4           | 96.7/97.7        | 95.2/97.2           |
| Screw   | 98.3/99.5     | 97.3/98.2           | 99.5/99.4        | 97.2/99.0           |
| Toothbrush | 100/88.6     | 99.9/99.3           | 100/99.9         | 99.6/99.8           |
| Transfer | 99.9/96.1     | 100/97.4            | 100/96.7         | 98.4/97.8           |
| Zipper  | 97.0/94.6     | 99.9/97.5           | 99.3/98.8        | 96.5/97.1           |
| Average | 97.6/43.4     | 96.4/97.2           | 99.1/97.9        | 94.6/97.3           |

Fig. 6. Quantitative results of parameters analysis on MVTec AD dataset. (a)–(d) AUROC [%] versus permutation of mutillevel features, output dimension $C_f$, early stop hyperparameter $\eta$, gradient operator $G$, respectively.

V. DISCUSSION

A. Parameters Analysis

In this section, we analyze five parameters introduced in our proposed method.

1) Selection of Pretrained Network $\phi$: In the first place, following previous representation based methods [21], [22], [28], [43], we conduct comparative studies on MVTec AD dataset with ResNet18 [48], Wide_ResNet50 [49], ResNeXt-101 [50] and Efficientnet_b4 [51] as the pretrained network $\phi$, and the results are shown in Table IV. It can be observed that the performance of anomaly detection improves as the network deepens. Moreover, the dimensions of the features extracted by the networks with different frameworks are significantly different, which has certain effects on the downstream tasks.

2) Permutation of Different Level Features: As shown in Fig. 6(a), leveraging only higher or lower-level features to represent images cannot achieve the best anomaly detection performance. This is because too much useful semantic information is lost. Therefore, the best results are obtained by using multiple levels of features (i.e., one–three level features).

3) Dimension of Feature Vectors $C_f$: The anomaly detection performance improves with the dimension of feature vectors in feature repository $M_f$ decreases, which can be observed in Fig. 6(b). This phenomenon proves that the multi-level aggregation features directly extracted by the pretrained network have redundancies in dimension, and removing these redundancies is beneficial for anomaly detection. However, a lower dimension of feature vectors does not always improve the results when the dimension $C_f$ less than 512.

4) Selection of Early Stop Hyperparameter $\eta$: A lower hyperparameter $\eta$ means a larger training epoch. As shown in Fig. 6(c), as the training epoch increases, the performance of our proposed method improves initially but then degrades rapidly, which justifies the need for the early stop strategy. Experimentally, the best result is achieved when $\eta = 0.8$.

5) Selection of Gradient Operators $G$: In our proposed method, some classical gradient operators [41] are considered for calculating the gradient score map $S_i$ that represents the degree of dramatic changes in multilevel aggregation feature $F_i$. Compared with the first-order differential gradient operators (e.g., Sobel operator, Robert operator, and Prewitt operator) that perform only one derivation, the second-order differential gradient operators (e.g., Laplacian operator) are more sensitive to dramatic changes in structure and texture by rederiving the gradients. In addition, the Canny operator can also capture dramatic changes in texture and structure, but it is not suitable for our proposed method on account of the loss of all smooth regions caused by the non-maximum suppression. Because the Laplacian operator has good rotation invariance and is easily implemented by convolution operations, we choose it as $G$ to calculate gradient score map $S_i$ in this work. Furthermore, the quantitative results with different gradient operators are reported in Fig. 6(d), showing our method with Laplacian operator achieves the best performance.

B. Ablation Study

We evaluate the contribution of each component of our proposed method by conducting a thorough ablation study and present the quantitative results in Table V. $P$, $G$, and $E$ modules represent feature extraction with pretrained network,
Moreover, using our proposed framework achieves encouraging results for both anomaly detection and localization. It can be shown that by combining the three modules we designed, our proposed framework achieves superior performance in anomaly detection and localization. In addition, our approach shows excellent superiority in few shot anomaly detection.

VI. CONCLUSION

In this article, we propose a novel discriminative feature learning framework with gradient preference for anomaly detection. By presenting gradient preference-based feature vector selection strategy and discriminative feature learning network with center constraint, the features extracted by the pretrained network can be brought close to better anomaly detection results. Extensive experiments on industrial and medical anomaly detection datasets demonstrate that our proposed method achieves competitive results in both anomaly detection and localization. In addition, our approach shows excellent superiority in few shot anomaly detection.

REFERENCES

[1] W. C. Li and D. M. Tsai, “Defect inspection in low-contrast LCD images using Hough transform-based nonstationary line detection,” IEEE Trans. Ind. Informat., vol. 7, no. 1, pp. 136–147, Feb. 2011.
[2] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, “PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection,” IEEE Trans. Ind. Informat., vol. 16, no. 12, pp. 7448–7458, Dec. 2019.
[3] K. Liu, N. Luo, A. Li, Y. Tian, H. Sajid, and H. Chen, “A new self-reference image decomposition algorithm for strip steel surface defect detection,” IEEE Trans. Instrum. Meas., vol. 69, no. 7, pp. 4732–4741, Jul. 2020.
[4] Q. Luo, X. Fang, L. Liu, C. Yang, and Y. Sun, “Automated visual defect detection for flat steel surface: A survey,” IEEE Trans. Instrum. Meas., vol. 69, no. 3, pp. 626–644, Mar. 2020.
[5] W. Xiao, K. Song, J. Liu, and Y. Yan, “Graph embedding and optimal transport for few-shot classification of metal surface defect,” IEEE Trans. Instrum. Meas., vol. 71, pp. 1–10, 2022.
[6] X. Cao, B. Chen, and W. He, “Unsupervised defect segmentation of magnetic tile based on attention enhanced flexible U-Net,” IEEE Trans. Instrum. Meas., vol. 71, pp. 1–10, 2022.
[7] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, “Towards MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection,” Int. J. Comput. Vis., vol. 129, pp. 1038–1059, Apr. 2021.
[8] Y. Huang, C. Qiu, and K. Yuan, “Surface defect saliency of magnetic tile,” Vis. Comput., vol. 36, no. 1, pp. 85–96, 2018.
[9] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 4183–4192.
[10] A. M. Kamoonah, A. K. Gostar, A. Bab-Hadiashar, and R. Hoseinnezhad, “Anomaly detection of defect using energy of point pattern features within random finite set framework,” 2021, arXiv:2108.12159.
[11] Z. Li et al., “Thoracic disease identification and localization with limited supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8290–8299.
[12] Y. Chen, Y. Tian, G. Pang, and G. Carneiro, “Deep one-class classification via interpolated Gaussian descriptor,” in Proc. AAAI Conf. Artif. Intell., 2022, vol. 36, no. 1, pp. 383–392.
[13] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, “f-AutoGAN: Fast unsupervised anomaly detection with generative adversarial networks,” Med. Image Anal., vol. 54, pp. 30–44, May 2019.
[14] P. Perera, R. Nallapati, and B. Xiang, “OOGAN: One-class novelty detection using GANs with constrained latent representations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2898–2906.
[15] 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, 2021.
[16] R. Chalapathy and S. Chawla, “Deep learning for anomaly detection: A survey,” 2019, arXiv:1901.03407.
[17] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, “GANomaly: Semi-supervised anomaly detection via adversarial training,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2018, pp. 622–637.
[18] D. T. Nguyen, Z. Lou, M. Klar, and T. Brox, “Anomaly detection with multiple-hypothesis predictions,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 4800–4809.
[19] E. Principi, D. Rossetti, S. Squartini, and F. Piazza, “Unsupervised electric motor fault detection by using deep autoencoders,” IEEE/CAA J. Autom. Sinca, vol. 6, no. 2, pp. 441–451, Mar. 2019.
