Position Encoding Enhanced Feature Mapping for Image Anomaly Detection

Qian Wan, Yunkang Cao, *Student Member, IEEE*, Liang Gao, *Senior Member, IEEE*, Weiming Shen, *Fellow, IEEE*, and Xinyu Li, *Member, IEEE*

***Abstract*—Image anomaly detection is an important stage for automatic visual inspection in intelligent manufacturing systems. The wide-ranging anomalies in images, such as various sizes, shapes, and colors, make automatic visual inspection challenging. Previous work on image anomaly detection has achieved significant advancements. However, there are still limitations in terms of detection performance and efficiency. In this paper, a novel Position Encoding enhanced Feature Mapping (PEFM) method is proposed to address the problem of image anomaly detection, detecting the anomalies by mapping a pair of pre-trained features embedded with position encodes. Experiment results show that the proposed PEFM achieves better performance and efficiency than the state-of-the-art methods on the MVTec AD dataset, an AUCROC of 98.30% and an AUCPRO of 95.52%, and achieves the AUCPRO of 94.0% on the MVTec 3D AD dataset.**

I. Introduction

Image anomaly is an image with parts of anomalous regions compared with the regular images [1]. Image anomaly detection is a task to figure the anomalies out, which is an important and effective stage for automatic visual inspection in intelligent manufacturing systems. The methods for image anomaly detection can help mitigate the long-time working of human labor and improve the efficiency of inspection.

In real-world industrial scenarios, the anomalies widely range over sizes, shapes, colors, etc. [2]. Some examples of anomalous images are shown in Fig. 1. It can be viewed that anomalies happen on different types of products during manufacturing or assembly. And the anomalies widely range. Furthermore, since the anomalous images are non-existent during the training stage, only normal images are used for training. This further increases the difficulty of image anomaly detection. To address the task of image anomaly detection, this paper proposes a novel Position Encoding enhanced Feature Mapping (PEFM) method. Some examples of scoring maps outputted by the proposed PEFM method are listed in Fig. 1. The scoring maps show that anomalies in images can be obviously detected.

Image anomaly detection has attracted a lot of attention recently [1]-[5]. Various recently developed methods can be categorized into image reconstruction-based and feature embedding-based. Because only normal images are available for training, image reconstruction by the convolutional auto-encoder neural network is always applied to capture the distribution of normal images and can be used for anomaly detection of the unseen images [3]-[10]. Though the testing image can be reconstructed, the anomalous regions in images would affect the reconstruction and lead to a low performance [8], [12], [14]. To avoid the problem of image reconstruction based methods, feature embedding is recently studied for image anomaly detection by using the nonlinear feature mapping of pre-trained convolutional neural networks [9]-[17]. The methods of this type usually fit the distribution of pre-trained features of normal images rather than the distribution of raw normal images. They have shown better performance on image anomaly detection. Recently CFLOW method [17] is proposed to capture the pre-trained feature distribution based on conditional normalizing flow to make anomaly detection, which shows high performance, but the efficiency requires an improvement.

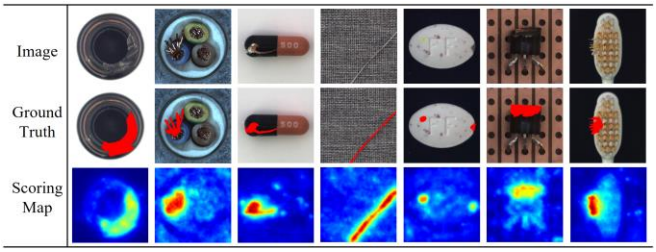

Fig. 1 The outputted scoring map of anomalous images by the proposed PEFM method. The images are from MVTec AD dataset [2]. Top rows are anomalous images, middle rows are ground truth marked in red color, and bottom rows are anomaly scoring maps.

In this paper, to improve the performance and efficiency of image anomaly detection, a novel Position Encoding enhanced Feature Mapping (PEFM) method based on feature embedding is proposed inspired by the previous work [12], [17], [19]. Considering the positional invariant of normal images to some extent, the position encoding proposed in [21] is used in CFLOW [19] to improve the detection performance. The proposed PEFM method encodes the pre-trained feature with position code and then maps it to another one pre-trained feature extracted by different pre-trained convolutional neural network. The proposed PEFM achieves a better performance and efficiency compared with the state-of-the-art (SOTA) methods. The main novelty and contribution in this paper are listed as follows:

Qian Wan, Yunkang Cao, Liang Gao, Weiming Shen, and Xinyu Li are with the State Key Laboratory of Digital Manufacturing Equipment & Technology, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail:wanqian19@hust.edu.cn;cyk_hust@hust.edu.cn;gaoliang@mail.hust.edu.cn;shenwm@hust.edu.cn;lixinyu@mail.hust.edu.cn,Corresponding author: *Xinyu Li*.)

(1) A novel Position Encoding enhanced Feature Mapping (PEFM) method is proposed in this paper to address image anomaly detection. (2) Experiments on the MVTec AD dataset are extensively conducted in this paper to show the improvement of the PEFM method.

The remaining of this article is organized as follows. The related work about image anomaly detection are stated in Section II. The proposed PEFM method is introduced in Section III. The experiments are conducted and analyzed in Section IV. Finally, Section V concludes this article.

II. Related work

The previous work on image anomaly detection can be categorized into image reconstruction-based and feature embedding-based and are stated in detail as follows.

A. Image Reconstruction based Methods

Because only normal images are available during the training stage, image reconstruction-based methods are studied to fit the distribution of normal images.

Structural similarity loss is applied to train a convolutional neural network to reconstruct the image [3]. The difference between the image and the reconstructed one is used for anomaly detection. The MemAE method [4] is proposed to learn a deep convolutional neural network with a bunch of normality in latent space to learn the normality of images in latent space. At the testing stage, the MemAE method applies the attention mechanism to design the new latent encodes to reconstruct the testing image, which can address the problem that anomalies in images can affect the reconstruction. The MemAE shows a good performance on anomaly detection compared with the vanilla auto-encoder. The MemAE is further extended for video anomaly detection [7]. A deep convolutional auto-encoder integrated with a generative adversarial network is applied to reconstruct the background of the original image. Specifically, a U-net-based network is trained to perform pixel-wise differences between the image and the reconstructed one for anomaly detection [8]. Iterative energy-based projection is used to augment the ability of reconstruction [9], which further improves the quality of the reconstructed image and shows a better performance compared to the vanilla auto-encoder. Artificial patches are embedded into training images to learn an inpainting convolutional neural network [10], improving the performance of anomaly detection.

Image reconstruction-based methods are intuitive and easy to understand, but the performance is limited due to the imperfect reconstruction.

B. Feature Embedding based Methods

To avoid the problem of image reconstruction based methods, feature embedding-based methods are further developed. Feature embedding-based methods make use of the deep convolutional neural network pre-trained on a large dataset, such as ImageNet [20], to embed images into nonlinear feature space, and detect anomalies in this space directly or indirectly.

Multivariate Gaussian is applied to fit parameters of MCNN distribution of pre-trained features of training normal images [12]. Then Mahalanobis distance is used as the anomaly score at the testing stage, showing a much better performance on image anomaly detection. The Padim [13] is proposed to fit the distribution of features in individual positions with a multivariate Gaussian, further improving the performance on image anomaly detection but increasing the number of parameters. Multiple independent multivariate Gaussian clustering is applied to represent the prototypes of normal images [14], and the minimum Mahalanobis distances between clusters are scores for anomaly detection. A pre-trained convolutional neural network is applied to calculate the difference between the image and reconstructed one as the score for anomaly detection [15]. Multiresolution knowledge distillation is designed to distill the knowledge from a pre-trained convolutional neural network to a raw one, and loss backpropagation is applied to detect the anomalies [16]. Knowledge distillation is directly used for image anomaly detection and achieves good performance and efficiency [17]. Normalizing flow is used to estimate the density of pre-trained features of training normal images and achieves a high performance [18]. The SOTA CFLOW method [19] is further proposed based on normalizing flow with a position condition.

Though the CFLOW method achieves high performance on the MVTec AD dataset, the efficiency requires further improvement.

III. The Proposed PEFM Method

In this paper, a novel Position Encoding enhanced Feature Mapping method is proposed to improve the performance and efficiency of image anomaly detection. As shown in Fig. 2, the framework of the proposed PEFM is introduced. The proposed PEFM method applies a pair of pre-trained convolutional neural networks to extract features. After that, the position code that is the same as in [19] is embedded with each pre-trained feature. Then encoded features are mapped to each other by a constructed mapping convolutional neural network. Minimizing mapping loss is adopted to train the mapping convolutional neural network during the training stage. And during the testing stage, the mapping residuals of testing images are fused as scores for anomaly detection. The details are introduced as follows.

A. Pre-trained Feature Extraction

Deep convolutional neural network like ResNet [22] pre-trained on a large image dataset has shown discriminative nonlinear feature extraction. In the proposed PEFM method, a pair of light convolutional neural networks is used for feature extraction.

Given an image $\mathbf{x}$, the pre-trained feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times D}$ of which are extracted by a pre-trained convolutional neural network (PCNN), denoted as follows:

$$\boldsymbol{F} = \mathrm{PCNN}(\mathbf{x}) \tag{1}$$

where H,W,D respectively denote the height, width and channels of the feature map. The pre-trained feature is normalized with L2-normalization. In the pre-trained feature space, the anomalies can be largely discriminated from normality, shown in previous feature embedding-based methods [12]-[14]. Besides, the multi-hierarchical pre-trained features can be applied to augment the detection of anomalies of different sizes.

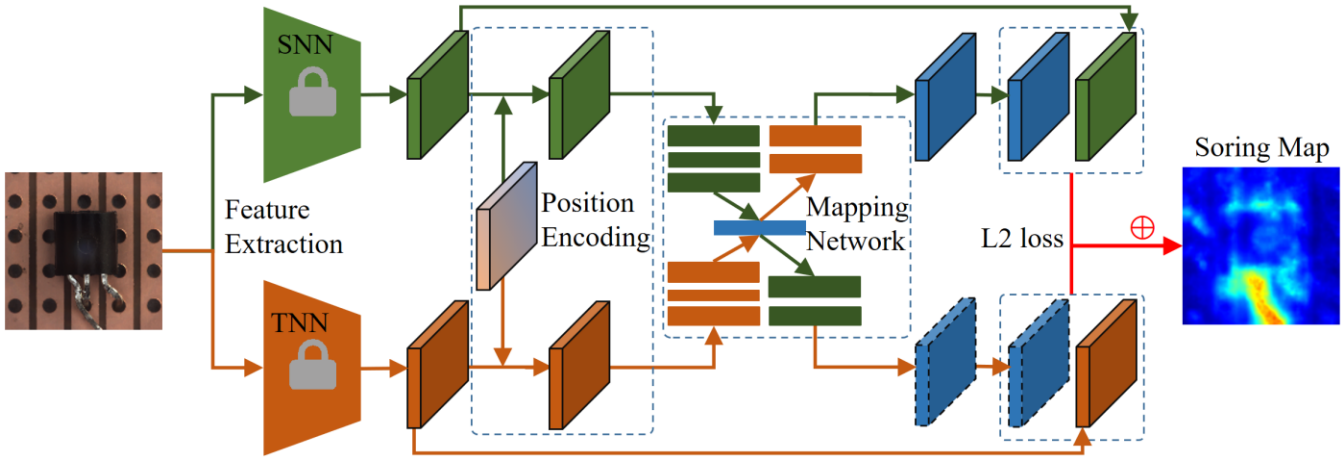

Fig. 2 The framework of the proposed PEFM method for image anomaly detection. The mapping convolutional neural network is applied to map position encoded pre-trained features to each other, and then the mapping residuals are fused as the score for anomaly detection.

B. Position Encoding

Considering there are some position invariants in normal images, the SOTA CFLOW method [19] takes the position codes [21] as the condition adding to features to estimate the density by the conditional normalizing flow. Inspired by this, the proposed PEFM method also takes the position invariant into considering, and encodes features in every position with the position code.

The position codes are a set of fixed values which are relative with position and require no training, the concrete format of 2-D position encode $\boldsymbol{E}$ is as follows:

$$\begin{aligned}
\boldsymbol{E}(x,y,2i) &= \sin(x/1000^{4i/D}) \\
\boldsymbol{E}(x,y,2i+1) &= \cos(x/1000^{4i/D}) \\
\boldsymbol{E}(x,y,2j+D/2) &= \sin(y/1000^{4j/D}) \\
\boldsymbol{E}(x,y,2j+1+D/2) &= \cos(y/1000^{4j/D})
\end{aligned} \tag{2}$$

where x, y respectively denote the axis position in height and width dimension of the feature map, D denotes the channels of feature map, and i, j respectively denote the axis position in channel dimension of feature map. Therefore, the position encodes of the whole feature map can be obtained through the Eq.(2).

C. Encoded Feature Mapping

In the proposed PEFM method, two pre-trained convolutional neural networks are used for feature extraction. The two networks respectively denote as source neural network (SNN) and target neural network (TNN), parameters of which are fixed during training stage. The corresponding feature maps $\boldsymbol{F}_S \in \mathbb{R}^{H\times W\times D_S}, \boldsymbol{F}_T \in \mathbb{R}^{H\times W\times D_T}$ are denoted as follows:

$$\boldsymbol{F}_S = \mathrm{SNN}(\mathbf{x}), \boldsymbol{F}_T = \mathrm{TNN}(\mathbf{x}) \tag{3}$$

where D_S, D_T respectively denote the number of channels of feature map.

To consider that the position invariant in training normal images to some extent, position encoding PE is applied to the extracted pre-trained features as follows:

$$\boldsymbol{EF}_S = \mathrm{PE}(\boldsymbol{F}_S, \boldsymbol{E}_S), \boldsymbol{EF}_T = \mathrm{PE}(\boldsymbol{F}_T, \boldsymbol{E}_T) \tag{4}$$

where $\boldsymbol{E}_S, \boldsymbol{E}_T$ respectively denote as the position encode of feature map $\boldsymbol{F}_S, \boldsymbol{F}_T$, and $\boldsymbol{EF}_S, \boldsymbol{EF}_T$ respectively denote as the encoded feature map. The type of PE can be achieved by adding or concatenation operation. The influence of PE is analyzed in section IV.

A mapping convolutional neural network (MCNN) is applied to map the encoded feature to the feature extracted by another pre-trained network. Given a set of training images, minimizing the mapping loss can be used to optimize the parameter of MCNN. The mapping residuals calculated for testing images can be scores for anomaly detection. The mapping is shown in Fig. 2, and can be denoted as follows:

$$\tilde{\boldsymbol{F}}_S, \tilde{\boldsymbol{F}}_T = \mathrm{MCNN}(\boldsymbol{EF}_S, \boldsymbol{EF}_T) \tag{5}$$

where $\tilde{\boldsymbol{F}}_S, \tilde{\boldsymbol{F}}_T$ respectively denote as the corresponding mapped feature map. The MCNN shares a part of parameters as shown in Fig. 2, to decrease the number of parameter. The optimizing loss function is design as following:

$$l = \frac{\left\|\tilde{\boldsymbol{F}}_S - \boldsymbol{F}_S\right\|_2^2 + \left\|\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_T\right\|_2^2}{2} \tag{6}$$

The parameters of MCNN are iteratively optimized by gradient decent in batch-wise.

After training, the mapping residual is used as the score for anomaly detection in the proposed PEFM method. The normal regions in the testing image have a much lower mapping residual, while the anomalous regions show a much higher mapping residual because the anomalous regions never be optimized during training. Therefore, the scoring map is calculate as following:

$$\boldsymbol{SM}(i,j) = \frac{\left\|\tilde{\boldsymbol{F}}_S(i,j) - \boldsymbol{F}_S(i,j)\right\|_2^2 + \left\|\tilde{\boldsymbol{F}}_T(i,j) - \boldsymbol{F}_T(i,j)\right\|_2^2}{2} \tag{7}$$

The multi-hierarchical features are applied in the proposed PEFM method to improve the detection performance because the different hierarchies capture different receptive fields. And the final scoring map is calculated as follows:

$$\boldsymbol{SM} = \frac{\sum_{k=1} \mathrm{resize}(\boldsymbol{SM}^k)}{K} \tag{8}$$

where K denotes the number of hierarchies, and $\text{resize}(\cdot)$ denotes the bilinear interpolation operation.

IV. Experiments

In this section, extensive experiments about the influence of position encoding, image size, and comparing with SOTA methods are conducted and analyzed.

A. Experimental Details and Settings

MVTec AD dataset [2] is an industrial dataset consisting of 15 categories for image anomaly detection. There are several hundred normal images in each category for training. The anomalies in testing images are widely ranged over sizes, shapes, colors, etc. All the testing images are provided with pixel-level ground truth.

MVTec 3D AD dataset [23] is also a real-world dataset consisting of 11 categories for image anomaly detection. There are several hundred normal images in each category for training. Only RGB data is used for training and testing in this paper.

TABLE I The Kernel size of each layer of MCNN.

Layer1	$1\times1\times c_S\times(c_S/2+c)$
Layer2	$1\times1\times(c_S/2+c)\times2c$
Layer3	$1\times1\times2c\times c$
Layer4	$1\times1\times c\times2c$
Layer5	$1\times1\times2c\times(c_T/2+c)$
Layer6	$1\times1\times(c_T/2+c)\times c_T$

Implementation Details. The images are normalized during the training and testing stage by the standard deviation and mean value of ImageNet dataset [20]. The other setting of images is the same as in [19]. The setting of SNN and TNN model are respectively set as pre-trained ResNet34 and ResNet50 convolutional neural network. The multi-hierarchical features are respectively extracted at the ReLU layer of the first three blocks of the ResNet. The MCNN model is designed with six layers which consist of one convolutional layer, one batch normalization layer, and one ReLU activation layer. The layer3 is shared in MCNN. And the kernel size of each layer is shown in TABLE I. The dimension of latent parameter is set as 200, 400, and 800 in three hierarchies. The setting of parameter c_S, c_T is relative with the type of position encoding. If the type of position encoding is adding operation PE_a in default, the setting as follows:

$$\begin{aligned} &c_S = D_S, c_T = D_T \\ &\boldsymbol{EF}_S = \boldsymbol{F}_S + \boldsymbol{E}_S, \boldsymbol{EF}_T = \boldsymbol{F}_T + \boldsymbol{E}_T \end{aligned} \tag{9}$$

And if the type of position encoding is adding operation PE_c, the setting as follows:

$$\begin{aligned} &c_S = 2D_S, c_T = 2D_T \\ &\boldsymbol{EF}_S = [\boldsymbol{F}_S, \boldsymbol{E}_S], \boldsymbol{EF}_T = [\boldsymbol{F}_T, \boldsymbol{E}_T] \end{aligned} \tag{10}$$

The MCNN is randomly initialized and optimized by the Adam with a learning rate of 3e-4, batch size of 16, weight decay of 1e-5, and the number of the training epochs is 200.

Evaluation Criterion. This paper follows the commonly used area under the receiver operating characteristic curve (AUCROC), and the normalized area under the per-region overlap curve (AUCPRO) when the false positive rate is lower than 0.3 [11] to evaluate the performance for pixel-level image anomaly detection. The commonly used frame per second (FPS) is adopted to compare the testing efficiency. The higher both criteria, the better performance of the method shows.

B. The influence of Position Encoding

The position encoding is important for ensuring the position invariant in normal images of some categories. Such as the Transistor category in the MVTec AD dataset shows a much strong position invariant. However, for the Screw, Tile, and Grid categories, the position variant is obvious between each image, so there is no need for insuring position invariant in these categories. Therefore, whether applying a position encoding is also experimentally analyzed. Two types of position encoding are compared.

TABLE II Exploring influence of Position Encoding in the proposed PEFM method. The experiment is conducted on MVTec AD dataset, and result is (AUCROC /%, AUCPRO /%).

Category	PEFM_n	PEFM_a	PEFM_c
Bottle	(98.48, 95.66)	(**98.51**, **95.92**)	(98.46, 95.76)
Cable	(97.03, 95.05)	(**98.31**, **97.74**)	(98.27, 97.29)
Capsule	(**98.67**, **93.31**)	(98.51, 92.11)	(98.56, 92.05)
Carpet	(**99.20**, **96.92**)	(99.15, 96.75)	(99.16, 96.76)
Grid	(**98.76**, **96.12**)	(98.46, 95.15)	(98.39, 94.60)
Hazelnut	(99.08, 96.67)	(**99.13**, 96.59)	(99.08, **96.70**)
Leather	(**99.38**, **98.72**)	(99.33, 98.60)	(99.35, 98.55)
Metal nut	(**97.49**, **95.06**)	(96.98, 93.88)	(97.08, 93.79)
Pill	(96.55, 95.50)	(97.04, **96.16**)	(**97.05**, 96.14)
Screw	(**99.03**, **95.11**)	(98.91, 94.52)	(99.00, 94.84)
Tile	(**96.02**, **88.19**)	(95.72, 87.52)	(95.77, 87.38)
Toothbrush	(98.65, 90.33)	(**98.74**, 90.81)	(98.72, **90.86**)
Transistor	(89.20, 76.51)	(96.78, **89.69**)	(**96.85**, 89.08)
Wood	(95.46, 92.35)	(**95.60**, 92.75)	(**95.60**, **93.01**)
Zipper	(98.37, 95.27)	(98.30, 95.04)	(**98.41**, **95.27**)
Average	(97.43, 93.38)	(97.96, **94.19**)	(**97.98**, 94.14)

As shown in TABLE II, the influence of position encoding is explored. PEFM_n denotes no position encoding, PEFM_a denotes that adding operation is the type of position encoding and PEFM_c denotes the concatenation operation. The average of the whole dataset shows that the position encoding achieves a better performance on image anomaly detection. Especially for Transistor and Cable two categories, there is much improvement because position invariant is obvious between images. For the Screw category, there is a drop in performance because the object in this category is rotating between images, so the position encoding cannot help detection anymore

TABLE III Exploring the influence of Image size in the proposed PEFM method. The experiment is conducted on the MVTec AD dataset, and result is (AUCROC /%, AUCPRO /%).

Category	128	256	512
Bottle	(98.06, 92.65)	(**98.51**, **95.92**)	(98.34, 96.30)
Cable	(98.09, 95.27)	(**98.31**, **97.74**)	(95.70, 93.59)
Capsule	(96.22, 77.20)	(**98.51**, 92.11)	(98.02, **95.75**)
Carpet	(98.53, 86.18)	(**99.15**, 96.75)	(99.05, **97.35**)
Grid	(94.63, 85.26)	(98.46, 95.15)	(**99.23**, **97.21**)
Hazelnut	(98.86, 89.12)	(99.13, 96.59)	(**99.17**, **97.99**)
Leather	(99.05, 97.51)	(99.33, 98.60)	(**99.42**, **98.91**)
Metal nut	(96.44, 88.61)	(**96.98**, 93.88)	(95.94, **94.63**)
Pill	(96.08, 92.02)	(**97.04**, **96.16**)	(96.10, 96.53)
Screw	(97.84, 90.65)	(98.91, 94.52)	(**99.01**, **95.73**)
Tile	(94.11, 77.58)	(95.72, 87.52)	(**96.55**, **91.10**)
Toothbrush	(97.99, 81.35)	(98.74, 90.81)	(**99.18**, **96.22**)
Transistor	(**98.39**, **90.84**)	(96.78, 89.69)	(90.39, 78.21)
Wood	(93.32, 79.55)	(95.60, 92.75)	(**96.49**, **95.77**)
Zipper	(96.24, 88.02)	(98.30, 95.04)	(**98.61**, **96.45**)
Average	(96.92, 87.45)	(**97.96**, 94.19)	(97.41, **94.78**)

From the TABLE II, two position encoding types achieve a better result. The concatenation operation increases the number of inputting channels to MCNN, increasing the number of parameters. Thus, adding operation is adopted in this paper since it carries no additional computation.

C. The influence of Image size

As pointed out in [19], the size of the inputting image can also affect anomaly detection performance. Thus, the influence of the image size is also discussed in this section. As shown in TABLE III, three scales of 128, 256, 512 have experimented. The 128 scale of the Transistor category performs the best result, also found in [19]. Besides, it can be viewed that 512 of most categories show a better performance for image anomaly detection. Therefore, the same setting of image size [19] is adopted in this paper.

D. Comparing with the SOTA methods on the MVTec AD and MVTec 3D AD datasets

The proposed method is compared with the strong baseline AE_{SSIM} [3], the recently proposed ST [11], STFPM [17], and the SOTA CFLOW [19] methods on the MVTec AD dataset. Also, the proposed PEFM is compared with AE [3], and STFPM [17] methods on the MVTec 3D AD dataset.

As shown in TABLE IV, the proposed PEFM method achieves a much better performance than the baseline method over 15 categories. The setting of image size is same as in CFLOW [19]. The PEFM is compared with the SOTA CFLOW method regarding the average performance on the MVTec AD dataset. From the average AUCPRO criterion, the PEFM method achieves better results. As shown in TABLE V, the proposed PEFM also achieves a comparing performance on the MVTec 3D AD dataset though it performs on some rotating object categories.

When comparing the testing efficiency, the size of images is fixed as 256. The testing environment is Intel i7 CPU and Nvidia GTX 2060 GPU. The proposed PEFM method achieves a much better result of 64.4 FPS, as shown in TABLE VI. Ingoring the data loading, the proposed PEFM is 2.7 times faster than the SOTA CFLOW method and 137.0 times faster than the PaDim method. This is because the two pre-trained convolutional neural networks used in the proposed PEFM method are small scale, and the whole testing can be achieved on a GPU device.

TABLE IV Comparing the proposed PEFM method with the baseline and SOTA methods on MVTec AD dataset. The result is (AUCROC /%, AUCPRO /%).

Category	AE_{SSIM}	ST	STFPM	CFLOW	$PEFM_a$
Bottle	(93.00, 83.40)	(-, 91.80)	(98.80, 95.10)	(**98.98**, **96.80**)	(98.51, 95.92)
Cable	(82.00, 47.80)	(-, 86.50)	(95.50, 87.70)	(97.64, 93.53)	(**98.31**, **97.73**)
Capsule	(94.00, 86.00)	(-, 91.60)	(98.30, 92.20)	(**98.98**, **93.40**)	(98.51, 92.11)
Carpet	(87.00, 64.70)	(-, 69.50)	(98.80, 95.80)	(**99.25**, **97.70**)	(99.15, 96.75)
Grid	(94.00, 84.90)	(-, 81.90)	(99.00, 96.60)	(98.99, 96.08)	(**99.23**, **97.21**)
Hazelnut	(97.00, 91.60)	(-, 93.70)	(98.50, 94.30)	(98.89, 96.68)	(**99.17**, **97.99**)
Leather	(78.00, 56.10)	(-, 81.90)	(99.30, 98.00)	(**99.66**, **99.35**)	(99.42, 98.91)
Metal nut	(89.00, 60.30)	(-, 89.50)	(97.60, **94.50**)	(**98.56**, 91.65)	(96.98, 93.88)
Pill	(91.00, 83.00)	(-, 93.50)	(97.80, **96.50**)	(**98.95**, 95.39)	(97.04, 96.18)
Screw	(96.00, 88.70)	(-, 92.80)	(98.30, 93.00)	(98.86, 95.30)	(**99.01**, **95.73**)
Tile	(59.00, 17.50)	(-, 91.20)	(97.40, 92.10)	(**98.01**, **94.34**)	(96.55, 91.10)
Toothbrush	(92.00, 78.40)	(-, 86.30)	(98.90, 92.20)	(98.93, 95.06)	(**99.18**, **96.21**)
Transistor	(90.00, 72.50)	(-, 70.10)	(82.50, 69.50)	(97.99, 81.40)	(**98.39**, **90.84**)
Wood	(73.00, 60.50)	(-, 72.50)	(**97.20**, 93.60)	(96.65, **95.79**)	(96.49, 95.77)
Zipper	(88.00, 66.50)	(-, 93.30)	(98.50, 95.20)	(**99.08**, **96.60**)	(98.61, 96.45)
Average	(87.00, 69.40)	(-, 85.70)	(97.0, 92.10)	(**98.62**, 94.60)	(98.30, **95.52**)

TABLE V Comparing the proposed PEFM method with the baseline and SOTA methods on MVTec 3D AD dataset. The result is (AUCPRO /%).

Category	AE	STFPM	$PEFM_n$	$PEFM_a$
Bagel	26.00	93.22	96.76	**97.52**
Cable gland	34.10	90.95	**98.46**	97.57
Carrot	58.10	97.48	**97.91**	97.72
Cookie	35.10	**91.73**	86.47	89.05
Dowel	50.20	89.80	**98.56**	98.02
Foam	23.40	69.82	**78.16**	77.34
Peach	35.10	93.30	96.73	**96.92**
Potato	65.80	95.49	96.52	**96.83**
Rope	1.50	94.54	**96.28**	95.09
Tire	18.50	88.41	**95.89**	94.83
Average	34.78	90.47	**94.17**	94.09

TABLE VI Comparing efficiency with SOTA methods.

FPS	PaDim	CFLOW	$PEFM_a$
Without dataloader	0.47	23.6	**64.4**
With dataloader	0.47	10.0	**14.4**

V. Conclusion

In this paper, a novel position encoding enhanced feature mapping method is proposed for image anomaly detection. The proposed PEFM method detects anomalies by mapping the position encoded feature to the feature extracted by another pre-trained convolutional neural network. It achieves a very good performance on the MVTec AD dataset with an AUCROC of 98.30% and an AUCPRO of 95.52%, and achieves the AUCPRO of 94.09% on the MVTec 3D AD dataset, and also shows a higher testing efficiency. The limitation of the proposed PEFM method is that it is sensitive to position invariants, and therefore an adaptive position encoding can be studied in future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 52188102.

References

[1] L. Ruff et al., "A Unifying Review of Deep and Shallow Anomaly Detection," in Proceedings of the IEEE, vol. 109, no. 5, pp. 756-795, May 2021.

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

[3] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders", Proceedings of the 14th International Joint Conference on Computer Vision Imaging and Computer Graphics Theory and Applications, 2019.

[4] Y. Cao, Q. Wan, W. Shen, and L. Gao, "Informative knowledge distillation for image anomaly segmentation," Knowledge-Based Systems, p. 108846, 2022.

[5] Q. Wan, L. Gao, L. Wang, and X. Li, "Partial Distillation of Deep Feature for Unsupervised Image Anomaly Detection and Segmentation," in Intelligent Computing Theories and Application, 2021, pp. 238–250.

[6] D. Gong et al., "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1705-1714.

[7] H. Park, J. Noh and B. Ham, "Learning Memory-Guided Normality for Anomaly Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14360-14369.

[8] C. Lv, F. Shen, Z. Zhang, D. Xu and Y. He, "A Novel Pixel-Wise Defect Inspection Method Based on Stable Background Reconstruction," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-13, 2021.

[9] D. Dehaene, O. Frigo, S. Combrexelle, and P. Eline, "Iterative energy-based projection on a normal data manifold for anomaly localization," in International Conference on Learning Representations (ICLR), 2020.

[10] V. Zavrtanik, M. Kristan, and D. Skcaj, "Reconstruction by inpainting for visual anomaly detection," Pattern Recognition, p. 107706, 2020.

[11] P. Bergmann, M. Fauser, D. Sattlegger and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4183-4192, Jun. 2020.

[12] O. Rippel, P. Mertens and D. Merhof, "Modeling the Distribution of Normal Data in Pre-Trained Deep Features for Anomaly Detection," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 6726-6733.

[13] T. Defard, A. Setkov, A. Loesch and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization", ICPR, 2020.

[14] Q. Wan, L. Gao, X. Li and L. Wen, "Industrial Image Anomaly Localization Based on Gaussian Clustering of Pretrained Feature," in IEEE Transactions on Industrial Electronics, vol. 69, no. 6, pp. 6182-6192, June 2022.

[15] D. S. Tan, Y. -C. Chen, T. P. -C. Chen and W. -C. Chen, "TrustMAE: A Noise-Resilient Defect Classification Framework using Memory-Augmented Auto-Encoders with Trust Regions," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 276-285, doi: 10.1109/WACV48630.2021.00032.

[16] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban and H. R. Rabiee, "Multiresolution Knowledge Distillation for Anomaly Detection," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14897-14907.

[17] G. Wang, S. Han, Errui Ding and D. Huang, "Student-teacher feature pyramid matching for unsupervised anomaly detection," Proceedings of the British Machine Vision Conference, 2021.

[18] M. Rudolph, B. Wandt and B. Rosenhahn, "Same Same But DifferNet: Semi-Supervised Defect Detection with Normalizing Flows," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1906-1915.

[19] D. Gudovskiy, S. Ishizaka and K. Kozuka, "CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1819-1828.

[20] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[21] A. Vaswani et al., "Attention is all you need," In Advances in Neural Information Processing Systems, 2017.

[22] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[23] P. Bergmann et al., "The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization," 17th International Conference on Computer Vision Theory and Applications, 2022.