Wavelet-SVDD: Anomaly Detection and Segmentation with Frequency Domain Attention

Linhui Zhou[1], Weiyu Guo[1(✉)], Jing Cao[2], Xinyue Zhang[1], and Yue Wang[1]

[1] School of Information, Central University of Finance and Economics, Beijing 102206, People's Republic of China
zhoulinhui@email.cufe.edu.cn, weiyu.guo@cufe.edu.cn

[2] China United Network Communications Group Co., Ltd., Beijing, China
caoj33@chinaunicom.cn

Abstract. Anomaly detection is a formidable challenge that entails the formulation of a model capable of detecting anomalous patterns in datasets, even when anomalous data points are absent. Traditional algorithms focused on learning knowledge regarding the typical features that arise in images, such as texture, shape, and color, to distinguish between normal and anomalous examples. However, there is untapped potential in frequency domain features for differentiating anomalous patterns, and current methodologies have not exhaustively exploited this avenue. In this work, we present an extension of the deep learning version of support vector data description (SVDD), a prevalent algorithm used for anomaly detection, through the introduction of Wavelet transformation and frequency domain attentions in the feature learning network. This extension allows for the consideration of frequency domain patterns in defect detection, and improves detection performance significantly. We performed extensive experiments on the MVTecAD dataset, and the results revealed that our approach attained advanced performance in both anomaly detection and segmentation localization, thereby confirming the efficacy of our proposed innovative designs.

Keywords: Anomaly detection · Wavelet transformation · Frequency domain attention

1 Introduction

Anomaly detection constitutes a pivotal binary classification issue that aims to detect the abnormalities in the data. This challenge persists across various industries such as finance, manufacturing, and video surveillance. Notably, a significant number of abnormal instances are either unattainable or inadequate for distribution modeling during training, anomaly detection is typically formulated as a semi-supervised or one-class classification task [8]. The identification of anomalies is particularly challenging in image data, as the difference between

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 230–243, 2023.
https://doi.org/10.1007/978-3-031-46664-9_16

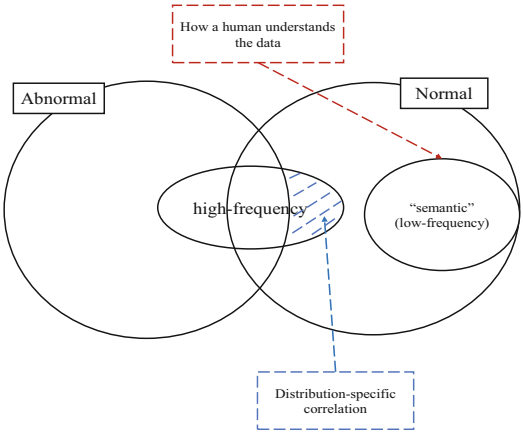

Fig. 1. Distribution of different frequencies

normal and anomalous patterns is often subtle, and defects can be nuanced, particularly in high-resolution images. Consequently, anomaly detection represents a distinctive binary classification problem that requires careful consideration, particularly in image data analysis.

Considering the diversity and scarcity of anomaly samples, a common strategy in such cases is to model the distribution of normal data and detect anomalies by identifying outliers. The pivotal aspect of this approach is to learn a concise boundary for normal data. In this regard, the support vector data description (SVDD) [12] and its extensions [10,15] have been employed as classical algorithms for one-class classification. These methods construct a data-enclosing hypersphere in the kernel space, enveloping most of the normal samples, for the purpose of anomaly detection. Nonetheless, existing works primarily focus on detecting semantic outliers, such as visual objects from distinct classes, in object-centric natural images, with little regard for the finer details, such as changes in texture, within an image. However, recent study [13] has illuminated that the features which can afford insight into the rate of transitions between pixels in an image are also useful to distinguish the abnormality from normal data. The potential for frequency domain features to effectively distinguish between normal and abnormal images deserves consideration.

As shown in Fig. 1, the low-frequency portion of an image is the primary source of semantic information perceived by the human visual system. This implies that, in anomaly detection tasks, the low-frequency features of normal images, which are consistent with the human visual system, share the same distribution, whereas its high-frequency features may not. However, tranditional Convolutional Neural Networks (CNNs) are probably capable of learning features that contain mixed high-frequency information [13], which may interfere with the construction of distribution of normal samples with the ability to distinguish abnormal samples by the deep SVDD model. In this work, we aim to

tackle the challenges related to the detection of image abnormalities and segmentation by means of integrating frequency domain features into CNNs. We present an innovative Wavelet attention that enables a more sophisticated distinction between normal and abnormal instances by incorporating frequency domain features. In this regard, a Wavelet Transform based network is proposed that extends the deep SVDD model to learn a precise boundary for normal data by considering both visual objects and frequency domain features. In a nutshell, our contributions in this work can be summarized as follows:

- We investigate the impact of frequency domain characteristics on the efficacy of anomaly detection, and put forward a multi-stage wavelet network that employs Wavelet attention to acquire knowledge pertaining to both the frequency domain features and visual objects, with the goal of improving image anomaly detection.
- We extend the classical method of Deep SVDD [10] for anomaly detection to frequency domain learning, and propose our Wavelet SVDD, which makes a good distinction between normal and anomalous in the feature space containing frequency domain features.
- A series of experiments are conducted to validate the effectiveness of the proposed method and the key designs, which demonstrate that our approach attained advanced performance in both anomaly detection and segmentation localization.

2 Related Works

This work aims to enhance the precision of anomaly detection through the incorporation of frequency domain feature learning into the framework of deep neural network-based Support Vector Data Description (SVDD). Its related work can be classified into three distinct categories: distance metric, frequency domain analysis, and frequency domain learning methods.

2.1 Distance Metric Based Methods

Distance-based methods focus on the training of a feature extractor that learns compact distribution of feature vectors derived from normal images by minimizing intra-class distances between samples. During the testing phase, the majority of methods employ the distance between the features of the sample undergoing evaluation and the normal features as a metric for detecting anomalies.

Deep support vector data description (Deep SVDD) [10] is a widely used technique in this domain. The authors of this approach artificially assign a point in the feature space as the feature center, and reduce the distance between the normal sample features and the center by mapping them to the center. Jihun et al. [15] expanded Deep SVDD to operate at the patch level by learning of the relative position semantics of patches through a self-supervised approach, thus avoiding the use of artificial centers by minimizing the distance between

semantically similar patches. However, notwithstanding the efficacy of the frequency domain feature analysis in detecting anomalies, existing distance-based methodologies have demonstrated a proclivity towards neglecting this avenue.

2.2 Frequency Domain Analysis Based Methods

The focus of frequency domain analysis based methods is on identifying irregularities in areas with regular textures. Previous methods [6,14] primarily rely on the manipulation of frequency spectrum information of the image, with the aim of removing periodic background textures and enhancing the visibility of anomalous regions. For example, Tianxiao et al. [14] involves the removal of frequency spectrum information of the background to highlight abnormal regions, while Chenlei et al. [6] employ only the phase spectrum to eliminate repetitive backgrounds in the inverse Fourier transform. However, these techniques have certain limitations in the case of image backgrounds and often require manual intervention for constructing periodic images. In contrast, our method learn the frequency domain features of the image, rather than relying solely on the spectrogram of the image.

2.3 Frequency Domain Learning Based Methods

The discrete Wavelet transform (DWT) [1] and Fourier transform (FT) [11] are widely employed image processing technique utilized for frequency domain analysis, which can transform an image from spatial domain to the frequency domain. Since the DWT can easily realize with the multi-level downsampling style, which is harmonious with deep convolutional neural networks (CNNs), it has been frequently combined with convolutional networks to deal with the tasks of computing vision.

For example, in order to enhance performance in the tasks of texture classification and image annotation, Shin et al. [5] proposed a wavelet-CNN architecture which incorporates a multiscale wavelet transform applied to the input image. This design has demonstrated superior performance compared to non-wavelet CNNs in these areas. Li et al. [9] presented an innovative solution to counteract the problem of feature loss encountered in wavelet-CNN [5]. They proposed to replace the downsampling features of CNNs with the low-frequency component of the discrete wavelet transform (DWT) and combining it with regular convolution, as opposed to spanwise convolution, resulting in improved feature retention. For a better fusion of spatial features and frequency domain features of the image, Zhao et al. [16] proposed an attention based network structure , i.e., Wavelet Attention (WA) block. The WA block first effectuates a decomposition of the feature map into low and high-frequency components through DWT's down-sampling operation. Subsequently, the high-frequency details of the feature map in the high-frequency component are selectively captured, while the essential information of the feature map residing in the low-frequency component remains undisturbed.

Previous research has demonstrated that high frequency information significantly impacts image classification, whereas we proposed a Wavalet Attention based SVDD approach to utilize an attention mechanism on the frequency domain to identify the relevant part of the high-frequency information for anomaly detection.

3 Methodology

Problem Formulation. The task of detecting anomalies is akin to binary classification in that it involves accurately distinguishing between normal and anomalous data. In the case of image anomaly detection, images that exhibit minor defects or those that fall outside the semantic distribution are typically deemed anomalous. To this end, various techniques have been proposed to learn a score function $\mathcal{A}_\theta$ to assess the level of anomaly in an image. Specifically, a high value of $\mathcal{A}_\theta(x)$ indicates that the image is anomalous during testing. Presently, the area under the receiver operating characteristic curve (AUROC) [3] is the standard metric employed to evaluate the efficacy of the $\mathcal{A}_\theta$ function in detecting anomalies, which is defined as:

$$AUROC(\mathcal{A}_\theta) = P(\mathcal{A}_\theta(X_{normal}) < \mathcal{A}_\theta(X_{abnormal})) \tag{1}$$

Ideally, an effective score function should be capable of assigning low and high scores to normal and anomalous input images, respectively. Moreover, for anomaly localization, the corresponding anomaly score is determined for each pixel.

Model Overview. As shown in Fig. 2, the proposed Wavelet SVDD involves two primary components: feature learning and anomaly calculation. Initially, the model employs a novel wavelet attention network to learn the feature distribution of normal images from both frequency domain features and visual objects. During the testing phase, we follow the pradigm of patch SVDD [15] to divide trained images into several patches and acquire the feature vectors of these patches by the Wavelet SVDD network, thereby enabling the separation of normal images into distinct patch distributions. Next, we extract the features of a testing image by the Wavelet SVDD network in manner of sliding windows. Finally, the distance between these extracted features from the testing image and the distribution of normal patches is treated as the abnormality score. The segmentation of abnormal pixels and the abnormality score of the entire image can be realized in manner of differentiation between testing image and the trained normal images.

3.1 Wavelet Transformation Network

The previous work [13] has demonstrated that the low-frequency portion of an image is the primary source of semantic information perceived by the human

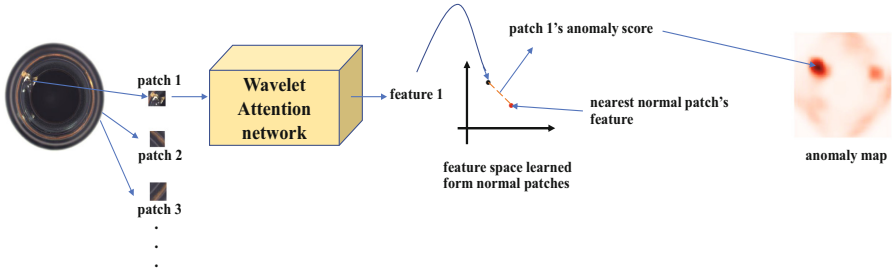

Fig. 2. Overall flow of the proposed model

visual system, and highlighted that high-frequency features are not extraneous noise; rather, a substantial number of them are correlated with the data distribution. Thus, we aim to enhance discrimination between normal and abnormal feature distributions by learning to filter out high-frequency information from the features obtained by CNNs, and select the effective high-frequency information from the filtered information.

The use of Discrete Wavelet Transform (DWT) in image processing has proven to be effective in obtaining high-quality down-sampling information while minimizing information loss in Convolutional Neural Networks (CNN). In this study, we aim to integrate the Discrete Wavelet Transform (DWT) into convolutional neural network (CNN) for frequency domain learning to enable the CNN to autonomously learn the components proficient in distinguishing anomalies among the frequency features generated by DWT. Specifically, we propose a Wavelet block which incorporates DWT operations into the feature extraction layers of CNN to enhance its performance. As illustrated in Fig. 3, we applies the DWT technique [1] with CNN to extract relevant features in the frequency domain. The Wavelet block first decomposes feature maps of CNN into low-frequency and high-frequency components by the DWT. The low-frequency component (X_{ll}) retains the primary information structure of feature maps, while the high-frequency components (X_{lh}, X_{hl}, and X_{hh}) store detailed information along with noise. Following the DWT, a 1×1 convolution layer and an Inverse Wavelet Transform (IWT) operation are stacked to select frequency features and convert them back into the spatial domain, respectively.

In line with previous work [9], the input of 2D-DWT $X \in R^{n \times n}$ can be obtained as follows:

$$\begin{aligned} \mathbf{X}_{ll} &= \mathbf{L}\mathbf{X}\mathbf{L}^{\mathrm{T}}, \ \mathbf{X}_{lh} = \mathbf{H}\mathbf{X}\mathbf{L}^{\mathrm{T}} \\ \mathbf{X}_{hl} &= \mathbf{L}\mathbf{X}\mathbf{H}^{\mathrm{T}}, \ \mathbf{X}_{hh} = \mathbf{H}\mathbf{X}\mathbf{H}^{\mathrm{T}} \end{aligned} \tag{2}$$

As a result of the biorthogonal property inherent in the Discrete Wavelet Transform (DWT), it is possible to reconstruct the original feature $\mathbf{X}$ with high accuracy and without any loss of information using the Inverse Wavelet Transform (IWT). The 2D-IWT is applied in accordance with the following procedure:

$$\mathbf{X} = \mathbf{L}^{\mathrm{T}}\mathbf{X}_{ll}\mathbf{L} + \mathbf{H}^{\mathrm{T}}\mathbf{X}_{lh}\mathbf{L} + \mathbf{L}^{\mathrm{T}}\mathbf{X}_{hl}\mathbf{H} + \mathbf{H}^{\mathrm{T}}\mathbf{X}_{hh}\mathbf{H} \tag{3}$$

where L and H are cyclic matrices composed of wavelet low-pass filter $\{l_k\}_{k\in Z}$ and high-pass filter $\{h_k\}_{k\in Z}$, respectively. Both these matrices have a size of $\lfloor N/2 \rfloor \times N$. L and H can be expanded as follows:

$$\mathbf{L} = \begin{pmatrix} \cdots & & & & \\ \cdots & l_0 & l_1 & \cdots & \\ & & \cdots & l_0 & l_1 & \cdots \\ & & & & \cdots & \cdots \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} \cdots & & & & \\ \cdots & h_0 & h_1 & \cdots & \\ & & \cdots & h_0 & h_1 & \cdots \\ & & & & \cdots & \cdots \end{pmatrix} \tag{4}$$

The Discrete Wavelet Transform (DWT) and Inverse Wavelet Transform (IWT) can be implemented as DWT and IWT layers in deep learning frameworks such as PyTorch, respectively. These layers operate on multichannel data on a per-channel basis. It should be noted that the wavelets chosen for use must possess finite filters to ensure that the size of the generated matrices is $\lfloor N/2 \rfloor \times N$. An example of a simple wavelet family is the Haar wavelet, which is characterized by a low-pass filter of $\{l_k\}_{k\in Z} = \{1/\sqrt{2}, 1/\sqrt{2}\}$ and a high-pass filter of $\{h_k\}_{k\in Z} = \{1/\sqrt{2}, -1/\sqrt{2}\}$.

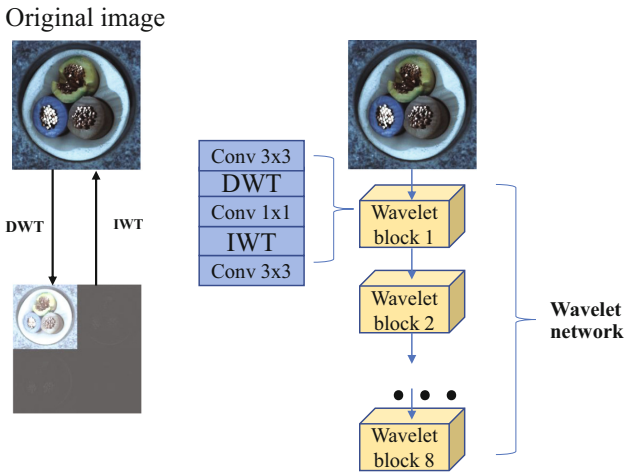

Fig. 3. Wavelet block

3.2 Wavelet Attention

In the presented network architecture, discerning and extracting valuable features from the frequency domain that are pertinent to anomaly detection has been successfully achieved. Nevertheless, Wang et al. [13] have observed that convolutional neural networks (CNNs) tend to prioritize learning low-frequency features in images. However, for anomaly detection, identifying high-frequency features, such as minute defects, is crucial in discriminating between anomalous

and normal instances. As a remedy, we further introduce an attention mechanism into the proposed Wavelet block to enable its CNN to concentrate more attention on the high-frequency elements.

Inspired by the Wavelet Attention mechanism proposed by Zhao et al. [16], we propose an enhanced Wavelet block based Wavelet Attention, which captures the detailed information of feature maps in the high-frequency component as the attention information, and the main information of feature maps in the low-frequency component $\mathbf{X}_{ll} = \{\mathbf{x}^{ll}\}_{i=1}^{N_p}$, is not affected. As Fig. 4 shown, the high-frequency components, i.e., $\mathbf{X}_{lh} = \{\mathbf{x}^{lh}\}_{i=1}^{N_p}$, $\mathbf{X}_{hl} = \{\mathbf{x}^{hl}\}_{i=1}^{N_p}$ and $\mathbf{X}_{hh} = \{\mathbf{x}^{hh}\}_{i=1}^{N_p}$, are selected and integrated into the low-frequency feature maps by an attention structure, which can be defined as:

$$\mathbf{z}_i = \mathbf{x}_i^{ll} + \frac{\exp(\mathbf{x}_i^{hl} + \mathbf{x}_i^{lh} + \mathbf{x}_i^{hh})}{\Sigma_{m=1}^{N_p} \exp(\mathbf{x}_m^{hl} + \mathbf{x}_m^{lh} + \mathbf{x}_m^{hh})} \mathbf{x}_i^{ll} \tag{5}$$

where $Np = H \times W$ is the number of elements on frequency feature maps.

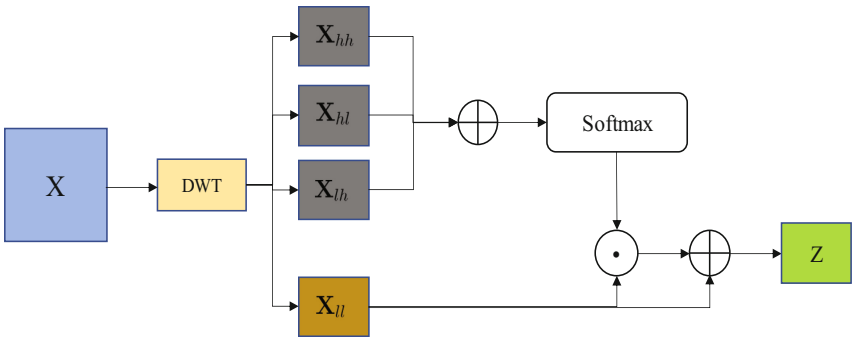

Fig. 4. Wavelet attention, $\oplus$ denotes broadcast element-wise addition, and $\odot$ denotes broadcast element-wise multiplication.

3.3 Wavelet SVDD

After analyzing the above information, we replaced the DWT decomposition part of our wavelet network with our wavelet attention to form our basic wavelet attention block. This basic wavelet attention block is then superimposed with the above wavelet network block to form a deep wavelet attention network as a feature learning network. We experimented with the depth and size of the network, as well as the number of wavelet attention network additions. Ultimately, we determined our feature learning network to be a network consisting of 4 layers of wavelet attention blocks and 4 layers of wavelet network blocks. Wavelet attention is placed in layers 2, 3, 6, and 7 after our experiments. The final network structure is shown in Fig. 5.

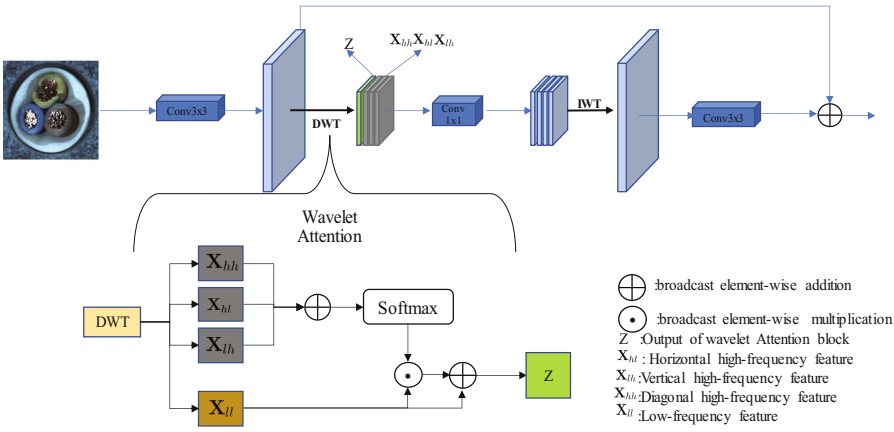

Fig. 5. Wavelet attention block

To enable our network to learn the distribution of normal images, we referred to the training method of patch svdd [15] and trained our network to collect semantically similar patches by itself. These semantically similar patches are obtained by sampling spatially adjacent patches. The encoder is then trained to minimize the distances between their features using the following loss function:

$$\mathcal{L}_{\text{SVDD}} = \sum_{i,i'} \|f_\theta(\mathbf{p}_i) - f_\theta(\mathbf{p}_{i'})\|_2 \tag{6}$$

where $\mathbf{p}_{i'}$ is a patch near $\mathbf{p}$ and f_θ is the wavelet attention network. Furthermore, to enforce the representation to capture the semantics of the patch and improve the structure of the anomalous and normal distributions, Wavelet SVDD appends the following self-supervised learning.

We followed the practice in patch SVDD based on Doersch et al. [4] and trained an encoder and classifier pair to predict the relative positions of two patches. A well-performing encoder pair means that the trained encoder can extract useful features for location prediction. For a randomly sampled patch $\mathbf{p}_1$, Doersch et al. [4] drew another patch $\mathbf{p}_2$ from a 3×3 grid in one of its 8 neighborhoods. If we let the true relative position be $y \in \{0, ..., 7\}$, the classifier C_ϕ is trained to correctly predict $y = C_\phi(f_\theta(\mathbf{p}_1), f_\theta(\mathbf{p}_2))$. We added a self-supervised learning signal by adding the following loss term:

$$\mathcal{L}_{\text{SSL}} = \text{Cross-entropy}\left(y, C_\phi\left(f_\theta\left(\mathbf{p}_1\right), f_\theta\left(\mathbf{p}_2\right)\right)\right) \tag{7}$$

As a result, the encoder is trained using a combination of two losses with the scaling hyperparameter λ, as presented in Eq. 8. This optimization is performed using stochastic gradient descent and the Adam optimizer [7].

$$\mathcal{L}_{\text{Wavelet}} \quad _{\text{Psvdd}} = \lambda \mathcal{L}_{\text{SVDD}} + \mathcal{L}_{\text{SSL}} \tag{8}$$

3.4 Calculate Anomaly Score

After training the feature learning network, the representations from the network are used to detect anomalies. First, the representation of every normal train patch [15], $f_\theta(\mathbf{p}_{\text{normal}})|\mathbf{p}_{\text{normal}})$, is calculated and stored. Given a query image x, for every patch p with a stride s within x, the L_2 distance to the nearest normal patch in the feature space is defined as its anomaly score using Eq. 9.

$$\mathcal{A}_\theta^{\text{patch}}(\mathbf{p}) \doteq \min_{\mathbf{p}_{\text{normal}}} \|f_\theta(\mathbf{p}) - f_\theta(\mathbf{p}_{\text{normal}})\|_2 \tag{9}$$

At the same time, to improve the stability of our method and avoid the appearance of query patches being affected by noise in the normal distribution, we also set another anomaly score calculation function $\mathcal{A}2_\theta^{\text{patch}}$. The difference between this and the above anomaly score calculation function is that $\mathcal{A}2_\theta^{\text{patch}}$ considers the next closest patches in addition to the closest patches to the query patches. This reduces the influence of noise in the training data to a certain extent. Therefore, $\mathcal{A}2_\theta^{\text{patch}}$ is defined as:

$$\mathcal{A}2_\theta^{\text{patch}}(\mathbf{p}) \doteq \frac{1}{2} \times \min_{\mathbf{p}_{\text{normal1}}\mathbf{p}_{\text{normal2}}} \|f_\theta(\mathbf{p}) - f_\theta(\mathbf{p}_{\text{normal1}}) - f_\theta(\mathbf{p}_{\text{normal2}})\|_2 \tag{10}$$

Patch-wise calculated anomaly scores are then distributed to the pixels. As a result, pixels receive the average anomaly scores of every patch to which they belong. We use $\mathcal{M}$ and $\mathcal{M}2$, calculated from the two scoring methods $\mathcal{A}$ and $\mathcal{A}2$, respectively, to represent the resulting anomaly maps.

We divided the size of 32 and 64 patches input into the network, respectively, to obtain different sizes of anomaly maps. We aggregate multiple maps using element-wise multiplication. The resulting anomaly map, M_{multi}, provides the answer to the problem of anomaly segmentation:

$$\begin{aligned} \mathcal{M}1_{\text{multi}} &\doteq \mathcal{M}1_{\text{small}} \odot \mathcal{M}1_{\text{big}} \\ \mathcal{M}2_{\text{multi}} &\doteq \mathcal{M}2_{\text{small}} \odot \mathcal{M}2_{\text{big}} \\ \mathcal{M}\text{blend}_{\text{multi}} &\doteq \mathcal{M}1_{\text{multi}} \odot \mathcal{M}2_{\text{multi}} \end{aligned} \tag{11}$$

where M_{small} and M_{big} are the generated anomaly maps with different scales of patches, respectively. The pixels with high anomaly scores in the map $M_{multi} = \{\mathcal{M}1_{\text{multi}}, \mathcal{M}2_{\text{multi}}, \mathcal{M}\text{blend}_{\text{multi}}\}$ are deemed to contain defects.

It is straightforward to address the problem of anomaly detection. The maximum anomaly score of the pixels in an image is its anomaly score, which can be expressed as:

$$\mathcal{A}_\theta^{\text{image}}(\mathbf{x}) \doteq \max_{i,j} \mathcal{M}_{\text{multi}}(\mathbf{x})_{ij} \tag{12}$$

4 Experiments

We selected the MVTecAD dataset [2] to test the effect of our improvements. This dataset consists of 15 classes of industrial images, each class categorized as either an object or texture. Ten object classes contain regularly positioned objects, while the texture classes contain repetitive patterns.

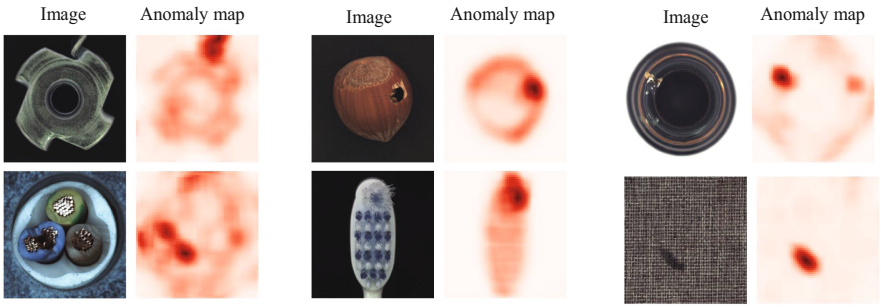

Fig. 6. Anomaly maps

Table 1. Detection and segmentation performance on MVTec AD

Classes	Det.	Seg.
bottle	0.996	0.987
cable	0.953	0.969
capsule	0.935	0.962
carpet	0.946	0.964
grid	0.949	0.965
hazelnut	0.964	0.978
leather	0.975	0.976
metal_nut	0.963	0.986
pill	0.946	0.965
screw	0.934	0.959
tile	0.984	0.941
toothbrush	1.000	0.983
transistor	0.943	0.969
wood	0.974	0.962
zipper	0.983	0.958
Average	0.963	0.968

Table 2. Detection and segmentation performance compared with baselines

Method	Det.	Seg.
InTra	0.950	0.966
PyramFlow	0.960	0.945
RegAD	0.927	0.966
CutPaste	0.961	0.883
Patch SVDD	0.921	0.951
Wavelet SVDD (Ours)	0.963	0.968

4.1 Anomaly Detection and Segmentation Results

Table 1 shows the detection performance of our method in each type of MVTecAD dataset in terms of AUROC. As shown in Fig. 6, the anomaly maps generated using the proposed method indicate that defects are properly localized, regardless of their size. Table 2 shows the detection and segmentation performances for the MVTecAD dataset compared with baselines in terms of AUROC.

4.2 Effect of Wavelet Attention

To explore the effect of our Wavelet attention block, we compared the performance of the network without the Wavelet attention block to that of the network with the Wavelet attention block added at different positions. Our network has mainly 8 wavelet layers. We compared the network without the Wavelet attention block to the network with the Wavelet attention block in layers 2, 3, 6, and 7, as well as in layers 3 and 6, in layers 2 and 7, and in all layers, respectively. The results on MVTec are shown in Fig. 7.

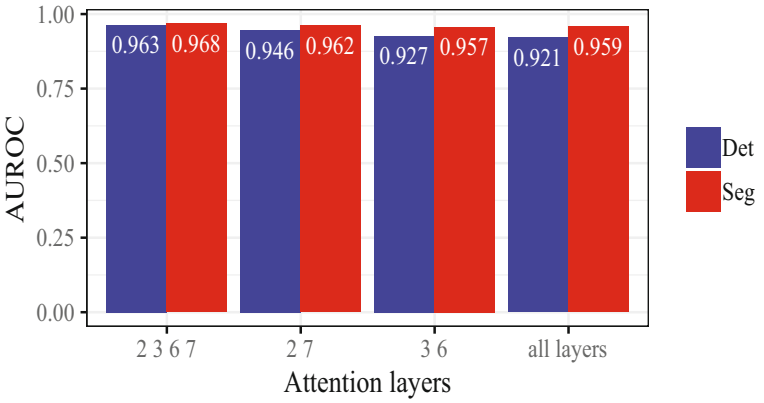

Fig. 7. Effects of different attention layers

The experimental results show that the accuracy of anomaly detection and segmentation can be improved by adding a Wavelet attention block in layers 2, 3, 6, and 7, respectively. The improvement of WA blocks in layers 2, 3, 6, and 7 is better than that in layers 2 and 7. In contrast, adding Wavelet attention blocks in all layers decreases accuracy. One possible explanation is that the influence of multiple Wavelet attentions creates a shortcut path dependence, which weakens the learning effect. Additionally, the frequency domain information in shallow layers may not be as useful for distinguishing anomalies as in deep layers. In conclusion, this experiment verifies the usability of Wavelet attention.

5 Conclusion

In this work, we present a novel technique for image anomaly detection and segmentation called Wavalet Attention SVDD. Instead of only relying on the conventional features extracted by convolutional network, we improve the patch SVDD [15] by involving the frequency domain characteristics of images to differentiate anomalies. We extensively evaluated our method on the MVTecAD dataset and observed that our approach outperformed existing techniques in

both anomaly detection and segmentation localization. These results validate the effectiveness of our innovative designs. However, the present approach inherits the inference architecture of patch SVDD, which necessitates anomaly detection inference based on feature database retrieval, thus resulting in time consumption. In future work, we plan to enhance detection inference by incorporating an Auto-Encoder structure into our detection model and accomplishing end-to-end learning and inference.

Acknowledgements. This work is jointly supported by National Natural Science Foundation of China(62106290) and Program for Innovation Research in Central University of Finance and Economics.

References

1. Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. IEEE Trans. Image Process. **1**(2), 205–220 (1992)
2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad-a comprehensive real-world dataset for unsupervised anomaly detection. In: ICCV, pp. 9592–9600 (2019)
3. Calders, T., Jaroszewicz, S.: Efficient AUC optimization for classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 42–53. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74976-9_8
4. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV, pp. 1422–1430 (2015)
5. Fujieda, S., Takayama, K., Hachisuka, T.: Wavelet convolutional neural networks. arXiv preprint arXiv:1805.08620 (2018)
6. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In: CVPR, pp. 1–8 (2008)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Li, C.-L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: self-supervised learning for anomaly detection and localization. In: CVPR, pp. 9664–9674 (2021)
9. Li, Q., Shen, L., Guo, S., Lai, Z.: Wavelet integrated CNNs for noise-robust image classification. In: CVPR, pp. 7245–7254 (2020)
10. Ruff, L., Vandermeulen, R., Goernitz, N., et al.: Deep one-class classification. In: ICML, pp. 4393–4402 (2018)
11. Tao, R., Zhao, X., Li, W., et al.: Hyperspectral anomaly detection by fractional Fourier entropy. IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. **12**(12), 4920–4929 (2019)
12. Tax, D.M., Duin, R.P.: Support vector data description. Mach. Learn. **54**, 45–66 (2004)
13. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: CVPR, pp. 8684–8694 (2020)

14. Wu, T., Wen, M., Wang, Y., et al.: Spectra-difference based anomaly-detection for infrared hyperspectral dim-moving-point-target detection. Infrared Phys. Technol. **128**, 104489 (2023)
15. Yi, J., Yoon, S.: Patch SVDD: patch-level SVDD for anomaly detection and segmentation. In: ACCV (2020)
16. Zhao, X., Huang, P., Shu, X.: Wavelet-attention CNN for image classification. Multimedia Syst. **28**(3), 915–924 (2022)