Hindawi
Journal of Sensors
Volume 2023, Article ID 6690190, 12 pages
https://doi.org/10.1155/2023/6690190

Research Article

Normalizing Flow-Based Industrial Complex Background Anomaly Detection

Pengxv Wen, Xiaorong Gao, Yong Wang, Jinlong Li, and Lin Luo

School of Physical Science and Technology, Southwest Jiaotong University, Chengdu 610031, China

Correspondence should be addressed to Xiaorong Gao; gxrr@vip.163.com

Received 5 August 2023; Revised 25 September 2023; Accepted 10 October 2023; Published 31 October 2023

Academic Editor: Rajkishor Kumar

This paper proposes a novel approach called cross-scale with attention normalizing flow (CSA-Flow) enhanced with channel-attention (CA) and self-attention (SA) modules for high-speed railway anomaly detection in complex industrial backgrounds to reduce the manual workload of the primary maintenance of high-speed electric multiple units. Detecting defects in industrial environments, characterized by intricate backgrounds and unclear subjects, poses significant challenges. To address this, CSA-Flow introduces a channel feature extraction module that combining the pretrained convolutional neural network models with a CA module for feature extraction, capturing information at different scales, and uses the SA module to capture more contextual information by its larger receptive field. The performance evaluation of CSA-Flow on the MVTec-AD dataset demonstrates an impressive area under the receiver operating characteristic curve (AUROC) score of 98.7%, with an equally remarkable score of 98.4% across all object classes. To further assess the effectiveness of CSA-Flow in complex background scenarios, we introduce a dedicated dataset, specifically designed for high-speed rail braking devices (HSRBDs). The experimental results establish the superiority of CSA-Flow over current state-of-the-art approaches in terms of both AUROC score and recall score, validating its exceptional capability for detecting anomalies in industrial complex backgrounds.

1. Introduction

Anomaly detection is a critical aspect within the field of railway detection. Safety is the foundation and primary concern of railway detection, directly impacting the lives of individuals and public property. Train accidents can be attributed to three primary factors: rail defects [1, 2], visual anomalies on the railway [3, 4], and misestimating or incorrect operation by the locomotive driver [4]. Any deviation from normal conditions is deemed an anomaly. In recent years, rapid advancements in image processing technology have drawn increasing attention to the detection of anomalies in the railway system. By prioritizing anomaly detection, we can ensure passenger well-being and safeguard critical public infrastructure.

In industrial applications, manual anomaly detection remains predominant in handling detection tasks. These approaches involve comparing visual texture features [5] between defective and normal samples to determine the presence of anomalies. However, manual detection methods often suffer from inefficiencies. As a result, deep-learning-based anomaly detection methods have gained traction in railway detection due to their inherent characteristics of speed, nondestructiveness, and high precision [6, 7]. Within the domain of high-speed railway, abnormal detection can be categorized into three main approaches: unsupervised methods [8], object detection methods [9–11], and defect segmentation methods [12, 13].

In real-world industrial detection, the scarcity of abnormal samples and the limited availability of labeled data present significant challenges for industrial anomaly detection. Additionally, in industrial applications, the backgrounds of detection objects are often complex, further compounded by the influence of moving parts, significantly raising the difficulty level of anomaly detection in these settings. To address these challenges, the MVTec-AD dataset [14] serves as a benchmark for anomaly detection, providing clear object boundaries where the previous methods have struggled to effectively incorporate contextual information. Consequently, our attention is directed toward exploring the self-attention

mechanism as a potential solution to enhance anomaly detection in industrial scenarios.

Given the complexities associated with complex backgrounds and unclear subjects in industrial anomaly detection, this paper aims to address practical challenges in this domain. During the training process, our focus lies in enabling the network to learn the distribution of normal samples only while differentiating between normal and abnormal samples during testing.

This approach is commonly referred to as semisupervised learning [15].

To tackle the aforementioned challenges, we compare related methods in Section 2, and normalizing flow (NF) methods demonstrate excellent anomaly location and industrial defect detection capabilities [16, 17], among others. We propose a semi-supervised anomaly detection method named cross-scale with attention normalizing flow (CSA-Flow), which utilizes NF [18, 19]. CSA-Flow specifically targets the problem of complex backgrounds and unclear subjects, allowing for the recognition and visualization of the defect regions within the image. It employs a full convolutional architecture and attention modules to establish global dependencies and expand the receptive field of the image. We evaluate the performance of CSA-Flow using the MVTec-AD dataset [14], designed to mimic real-world industrial inspection scenarios, as well as the BeanTech Anomaly Detection (BTAD) dataset [20]. Our proposed method achieves state-of-the-art accuracy in abnormality detection. Additionally, we apply CSA-Flow to a real high-speed rail braking device (HSRBD) dataset, which is one of the key components on the train, demonstrating its effectiveness in achieving high performance in real-world industrial applications.

The contributions of this paper are outlined as follows:

(1) Proposal of CSA-Flow incorporating the channel-attention (CA) module and the self-attention (SA) module to enhance anomaly detection accuracy by effectively capturing key features from input images.
(2) Achievement of state-of-the-art accuracy demonstrated by CSA-Flow on the MVTec-AD dataset and the BTAD dataset.
(3) Establishment of a real HSRBD dataset with complex backgrounds for anomaly detection and achieves a state-of-the-art accuracy.

2. Related Work

2.1. Reconstruction-Based Methods. Reconstructed image anomaly detection is a widely employed unsupervised approach for anomaly detection. The fundamental principle underlying this method is to model normal data in order to identify abnormal data that deviates from the learned model [21]. The core framework of this approach involves training a generative model using normal datasets and subsequently employing the model to reconstruct unseen data. By establishing a threshold for reconstruction error, any reconstructed data surpassing this threshold is considered anomalous. This methodology allows for the identification of anomalies based on deviations from the expected reconstruction patterns.

Autoencoder (AE) [22] is a widely utilized technique for anomaly detection, relying on the principle of reconstruction. AE is a type of neural network that compresses input data into lower-dimensional latent space and subsequently reconstructs it back to its original form by a decoder. In AEs architecture, the encoder processes the input data, extracting meaningful features and encoding them into a compressed representation. The decoder then decodes it back, reconstructing the data to resemble the original input. During the training phase, the AE aims to minimize reconstruction errors, ensuring that the output data closely matches the input data.

Similar to the decoding part of AE, the generator in generative adversarial networks (GANs) can be used for anomaly detection. Rudolph et al. [23] proposed to learn an inverse generator after training GAN and use both for reconstruction and error consideration.

Schlegl et al. [21] introduced AnoGAN, which aims to learn the manifolds of normal images from potential spaces, enabling the identification of anomalies in new images. Zenati et al. [24] trained a BiGAN model that simultaneously maps the image space to the latent space, showcasing improved statistics and computational outcomes. Akcay et al. [25] proposed GANomaly, building upon the concept of training GANs to learn the distribution of normal data and subsequently reconstructing input data using GANs. These innovative approaches leverage the power of deep learning and generative models to detect anomalies by learning normal data patterns and effectively reconstructing input data.

2.2. Embedding Similarity-Based Methods. These methods employ deep neural networks to extract meaningful vectors [26] or image blocks [27] to effectively describe the entire image for anomaly detection. Cho et al. [17] introduced a method known as semantic pyramid anomaly detection (SPADE). SPADE utilizes k-nearest neighbor (kNN) methods and leverages deep pretrained features. The proposed method focuses on aligning abnormal images with a series of similar normal images. SPADE introduces a novel approach that utilizes a multiresolution feature pyramid, allowing for a comprehensive analysis of image features across different scales.

Defard et al. [28] introduced PaDiM, a method that leverages pretrained convolutional neural networks (CNNs) for patch embedding. PaDiM uses multiple Gaussian distributions to generate a probability representation of normal data and employs the correlation among different semantic layers of the CNN to accurately identify the location of defects.

These methods employ the extraction of nominal features from the pretrained backbone networks, which are then utilized to construct a memory bank. During testing, the features extracted from the test images are compared against the entries in the memory bank. One significant advantage of this approach is its rapid speed, as the memory bank is preserved during training, requiring only feature

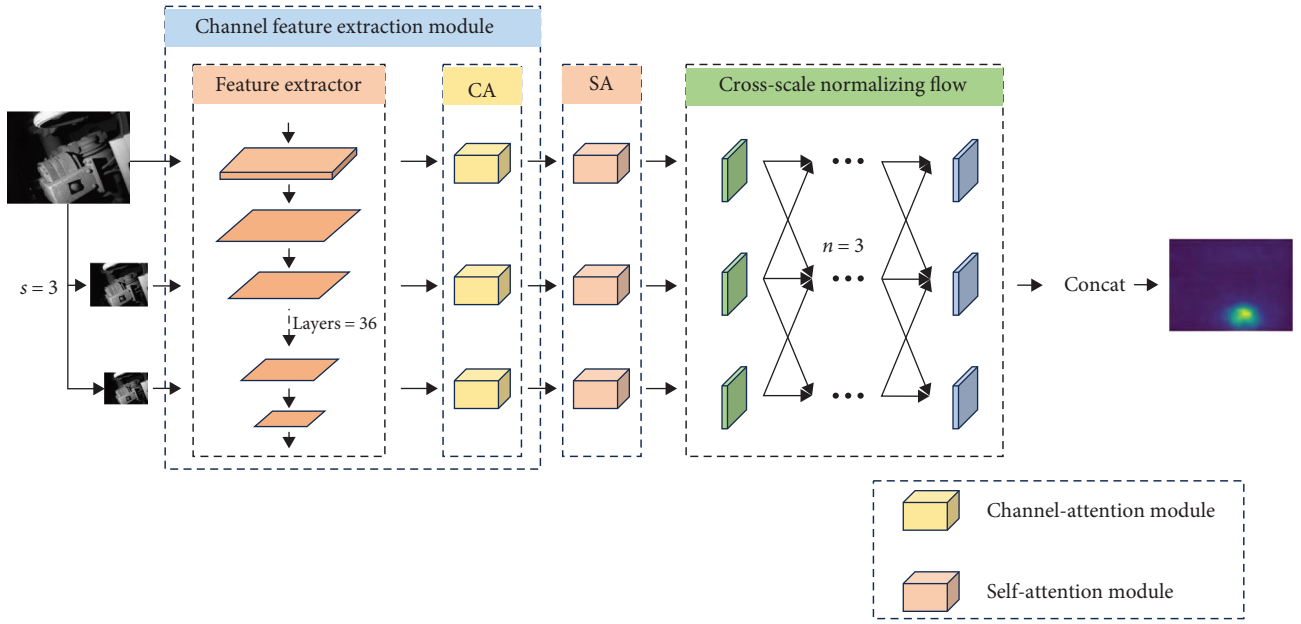

FIGURE 1: The pipeline of CSA-Flow.

comparisons during testing. However, some notable drawbacks are that the images stored in the memory bank must exhibit a high level of alignment and might not perform as well as other methods on large datasets.

2.3. NFs. NF is a distinctive generative model that sets itself apart from other models by its capability to generate distributions that are easily manageable. This feature enables efficient and accurate sampling as well as density evaluation. NF achieves this by employing reversible and differentiable mappings to transform a simple probability distribution, such as a normal distribution, into a more complex one [29]. In the NF framework, the density of a sample is converted back to the original sample distribution. The density evaluation of the sample involves calculating the product of the transformed sample's density and the volume change induced by the transformation. According to the change of variable formula, the volume change is determined by the absolute value of the Jacobian determinant at each transformation. NICE [18] and Real-NVP [30] are two notable examples of classic NFs that possess high speed in both forward and reverse processes. There are still some limitations in NF, especially when the distribution of abnormal data is very similar to the distribution of normal data, which can produce false positives.

In the field of anomaly detection, DifferNet [19] employs the NF estimation method to perform accurate likelihood tests, resulting in effective anomaly detection at the image level. However, due to the flattening of the output in DifferNet, it fails to locate the specific anomaly regions within the detected defects. To address this limitation, Gudovskiy et al. [31] introduced CFlow, which utilizes a discriminant pretraining encoder followed by a multiscale-generating decoder. This architecture allows for explicit judgment of the probability of encoding features. However, its effectiveness may vary when applied to more complex datasets.

3. Method

The proposed method, called CSA-Flow, is built upon the foundation of CS-Flow [32], a cross-scale normalized flow approach. CSA-Flow integrates the CA module and SA module to enhance the accuracy on common and realistic datasets while maintaining the high performance achieved by CS-Flow. Figure 1 provides an overview of the proposed method, illustrating its key components and workflow.

Similar to DifferNet [19], our approach initially involves training a model to learn features $y \in Y$ from defect-free images $x \in X$, enabling the detection of anomalies. During the evaluation process, we utilize density estimation of the extracted feature y to assign a similarity measure to each image x. A lower similarity score indicates a higher likelihood of an anomaly being present. Density estimation is achieved through bijective mapping, which involves learning from the unknown distribution p_Y in the feature space Y to the Gaussian distribution p_Z in the potential space Z. By leveraging the bidirectional mapping capability of NF, we utilize density estimation to map from the unknown distribution p_Y in the feature space Y to the Gaussian distribution p_Z in the latent space Z. Figure 1 illustrates the pipeline of CSA-Flow, depicting the various stages and transformations involved in the process.

For each category, we begin by computing the receiver operating characteristic (ROC) curve and identifying the optimal threshold θ, which maximizes the ratio of true positive rate (TPR) to false positive rate (FPR). Utilizing this selected threshold θ, we can determine whether a test image is abnormal or not:

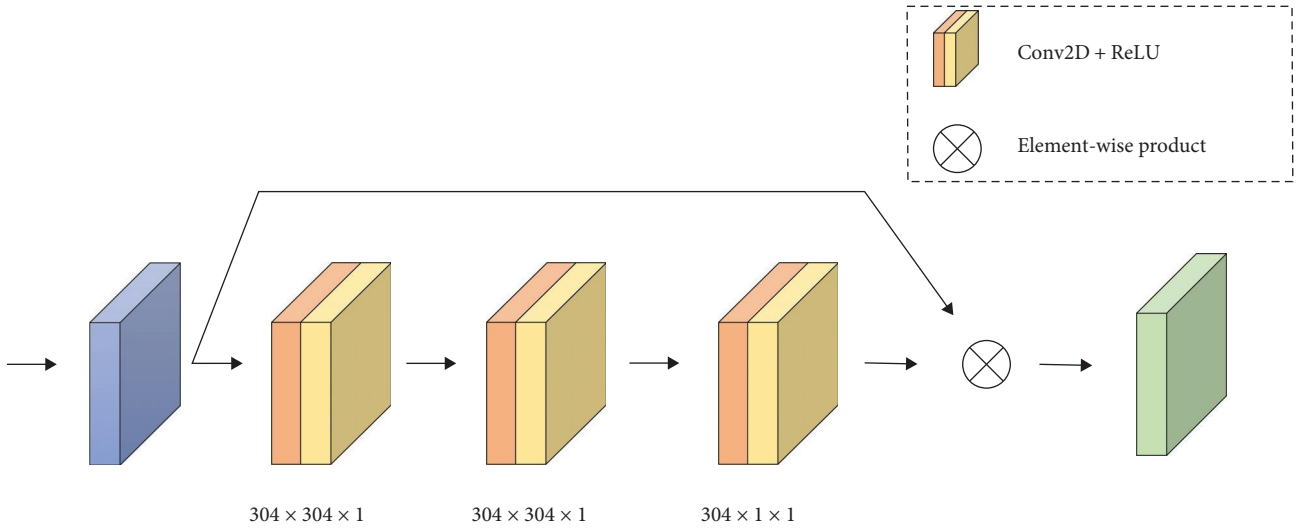

Figure 2: Channel attention (CA). Here, the symbol ⊗ denotes the element-wise product.

$$A(x) = \begin{cases} 1\, p_z(z) > 0 \\ 0 \text{ else} \end{cases}. \tag{1}$$

3.1. Channel Feature Extraction Module. Bergman and Hoshen [26] have demonstrated the exceptional performance of the ImageNet training feature extraction model for anomaly detection. Hence, we adopt feature extraction utilizing EfficientNets. The pretrained CNN possesses the capability to provide relevant features for anomaly detection [33]. Consequently, we employ a CNN that has been pretrained on ImageNet to extract the features y from the input image x. To enhance the descriptive capacity of the feature maps, we conduct a feature extraction on s images with varying resolutions. Subsequently, the images are segmented into multiple scales, leveraging techniques such as upsampling and stride convolutions to adjust the input image scale. The NF architecture excels in performing intensive data estimation, enabling it to effectively preserve detailed location and context information.

The first subnetwork is the channel feature extraction module combining the CA module [34] with CNNs feature extraction. It leverages scalar values to represent and evaluate the significance of each channel in an image. Let's assume $X \in \mathbb{R}^{H \times W \times C}$ is the image feature tensor in the network, where C is the number of channels, H is the feature height, and W is the feature width [35]. Figure 2 illustrates the architecture of CA. The prediction is generated using the following formula:

$$F' = M_c(F) \otimes F, \tag{2}$$

where F represents the input of CA and $M_c(\cdot)$ corresponds to the CA module.

By leveraging the channel feature extraction module, we believe that the model becomes more adept at focusing on valuable information.

3.2. Cross-Scale Flow. The cross-scale NF method proves to be highly effective in image anomaly detection. It processes feature maps at different scales to capture diverse information, leveraging the interplay between these scales to share relevant insights. Furthermore, the module's fully convolutional nature ensures the spatial dimensions are preserved, enabling accurate localization of anomalies. The cross-scale flow consists of a series of affine transformations implemented through coupling blocks. Based on the reference to the coupling blocks described in the study of Dinh et al. [30], we adopt the basis architecture of Real-NVP, as illustrated in Figure 3. The network estimates each scale and offset coefficient estimated by the subnetworks, denoted as r_1 and r_2, so that each input tensor y^i_{in} is randomly divided into $y^i_{in,1}$ and $y^i_{in,2}$. The obtained parameters are then employed as shown:

$$y_{\text{out},2} = y_{in,2} \odot \exp\left(\gamma_1 s_1\left(y_{\text{in},1}\right)\right) + \gamma_1 t_1\left(y_{\text{in},1}\right), \tag{3}$$

$$y_{\text{out},1} = y_{in,1} \odot \exp\left(\gamma_1 s_2\left(y_{\text{in},2}\right)\right) + \gamma_2 t_2\left(y_{\text{in},2}\right), \tag{4}$$

where the symbol $\odot$ denotes the element-product operation.

3.3. SA Module. The neural network processes a vast amount of vectors with varying sizes and connects them. However, this approach may not effectively uncover the intrinsic relationships among the inputs during training, resulting in suboptimal learning outcomes. To address this limitation, CSA-Flow incorporates an SA module, as depicted in Figure 4, to emphasize the correlations between features at different scales. The self-attention mechanism enables the model to establish global dependencies and expand the receptive field of an image. Compared to CNN, the SA module has a larger receptive field, allowing it to capture more contextual information.

The attention module can be represented by a set of queries and key-value pairs. The output is computed as a

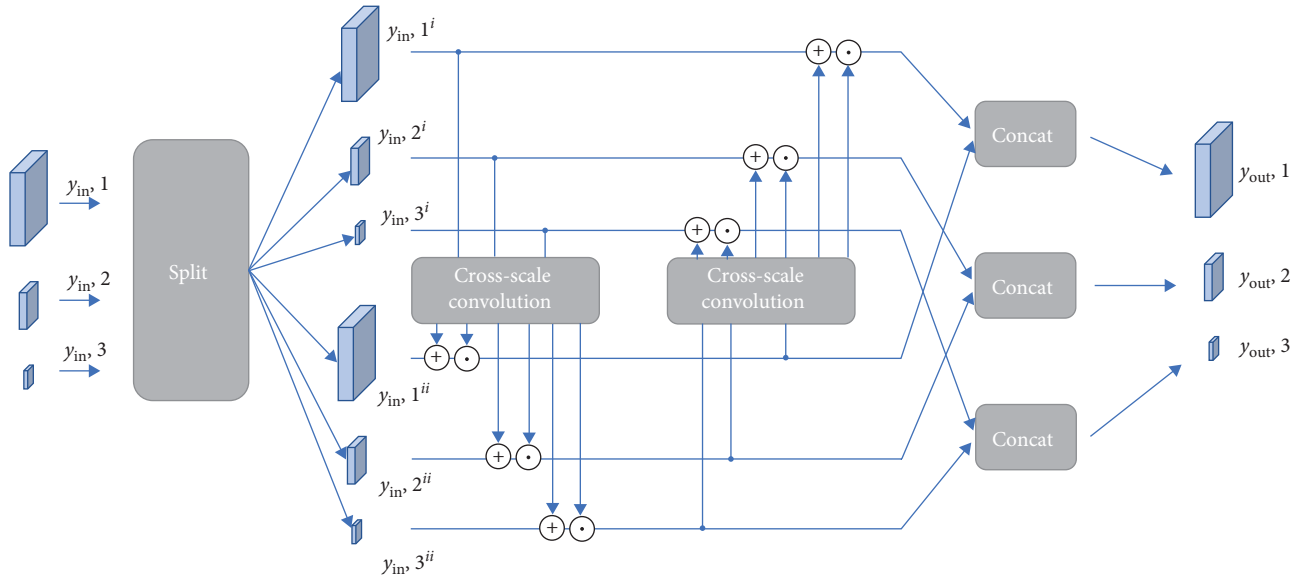

FIGURE 3: The structure of the cross-scale convolution is based on Real-NVP.

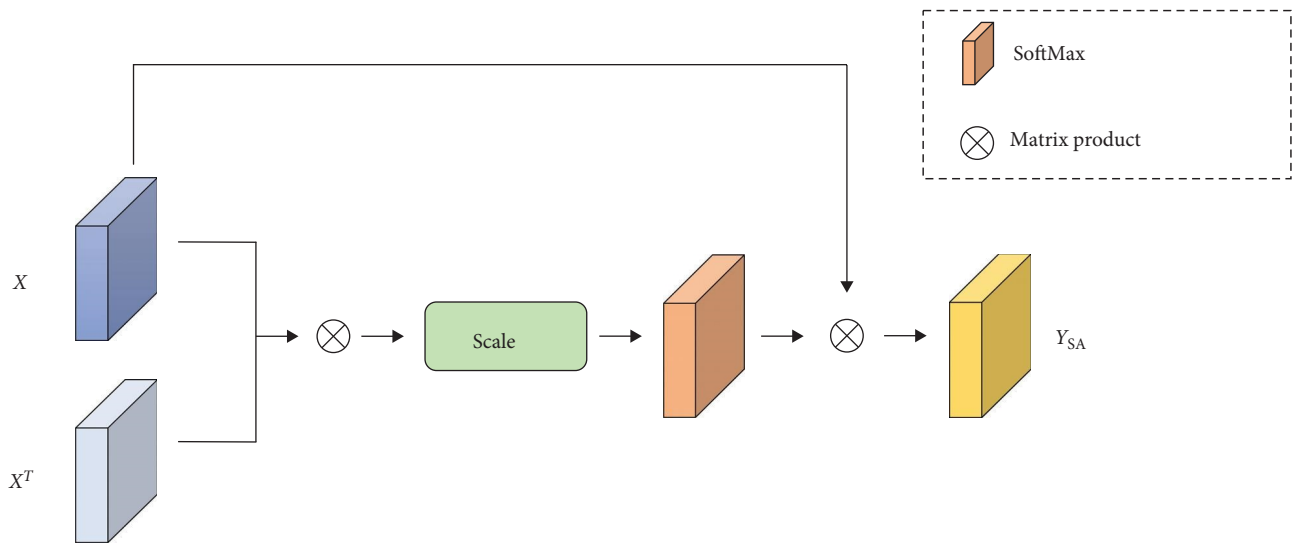

FIGURE 4: The structure of the self-attention module.

weighted sum of values, where each weight is determined by the correlation between the query and the key [36]. The Y_{SA} representation, which captures the relationship between pixels by using the dot-product of the query and the key as the weight, is formulated as follows:

$$Y_{SA} = \text{softmax}\left(\frac{XX^T}{\sqrt{d_k}}\right)X, \tag{5}$$

where X represents the extracted feature, Y_{SA} represents the feature map that contains the information necessary for detecting co-occurrence relationship anomalies, and d_k represents the depth of X [36].

3.4. Negative Log-Likelihood Loss. The objective of the training process is to maximize the likelihood of the mapping from the latent space Z to the feature space X. We adopt the likelihood formulation proposed in the study of Rudolph et al. [32] as follows:

$$p_Y(y) = p_Z(z)\left|\det\frac{\partial z}{\partial y}\right|, \tag{6}$$

which aims to maximize the log-likelihood. Similar to the study of Rudolph et al. [32], we utilize the negative log-likelihood loss $\mathscr{L}(y)$ to train the proposed model as follows:

$$\log p_Y(y) = \log p_Z(z) + \log\left|\det\frac{\partial z}{\partial y}\right|, \tag{7}$$

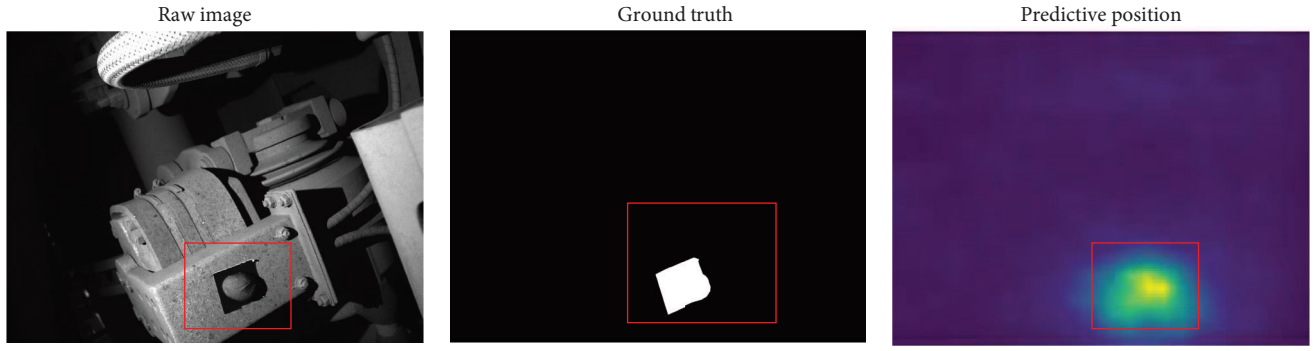

FIGURE 5: Demonstration of defects in industrial datasets (HSRBD). The one on the left (a) is the anomaly picture, the one in the middle, (b) is the ground truth, and the one on the right, and (c) is the predictive positioning.

$$\mathscr{L}(y) = \frac{\|z\|_2^2}{2} - \log\left|\det\frac{\partial z}{\partial y}\right|, \tag{8}$$

where $\|z\|_2^2$ represents the squared l_2-norm of a vector x in n-dimensional Euclidean space, which is defined as the sum of the squares of its components. The term $\det\frac{\partial z}{\partial y}$ represents the absolute determinant of the Jacobian. To ensure stability, we constrain the gradients of the l_2-norm to be equal to one.

4. Experiments

4.1. Datasets and Metrics. We evaluate the proposed method in various defect detection scenarios using the MVTec anomaly detection (MVTec-AD) dataset [6]. The MVTec-AD dataset, introduced by Bergmann et al. [14, 27], is designed to simulate anomaly detection in industrial applications. It offers high-resolution images with variations in multiple scales and lighting conditions. The dataset consists of 15 classes, including 10 object classes and 5 texture classes, each containing both normal and abnormal samples. The training set exclusively comprises defects-free images, while the test set consists of normal and abnormal images.

The BTAD dataset [20] includes 2,540 images of three industrial product categories. The training sets exclusively include normal samples, while testing sets contain both normal and abnormal samples. To assess the performance of the proposed method in real industrial applications, we curated a real-world dataset called the HSRBD dataset. This dataset comprises four scenarios that represent real-world HSRBDs. Each scenario includes four different industrial components with unknown size and foreign matters. Within each scenario, there are a varying number of high-resolution images, ranging from 160 to 220, with dimensions of $2{,}064 \times 1{,}544$ pixels. The presence of dynamic lighting and moving parts in each scenario adds complexity to the anomaly detection task, making it more closely aligned with actual application scenarios, as illustrated in Figure 5.

To evaluate the performance of the proposed method, we compute the area under the receiver operating characteristic curve (AUROC) on the publicly available datasets MVTec-AD and BTAD. In industrial applications, a more intuitive metric is needed. Therefore, we also compute the recall, which represents the detection rate of anomalies on the real-world HSRBD datasets. The TPR and FPR are defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \tag{9}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{10}$$

AUROC reflects the classifier performance by measuring the AUROC [37]. The classifier with a larger AUROC value indicates a better accuracy of the classifier. On the other hand, the recall rate, also known as the detection rate, measures the proportion of positive cases correctly identified by the classifier. The recall rate is a measure of coverage and is equivalent to sensitivity.

4.2. Implementation Details. To achieve a balanced combination of feature semantic level and spatial resolution, we utilize the output of the 36th layer of the pretrained EfficientNet-*B5* model from ImageNet as the feature extractor in all our experiments. For the CA module, we set the input channel and the output channel sizes to 304 to match the dimensionality of the extracted features. To standardize the input image size, we resize the images to $1{,}024 \times 1{,}024$ for the real-world HSRBD datasets. We extract features at three different scales: $(1{,}024 \times 1{,}024)$, (512×512), and (256×256). For the MVTec-AD dataset, we resize the input images to 768×768. In our implementation, we employ four coupling blocks ($n_{\mathrm{blocks}} = 4$). The internal networks of the first three blocks use 3×3 convolutional kernels, while a 5×5 convolutional kernel is applied in the last block. We set the negative slope of leaky ReLU to 0.1 and set the clamp parameter to $\alpha = 3$. For optimization, we utilize the Adam algorithm [38] with a learning rate of 2×10^{-4}, weight attenuation of 10^{-5}, momentum value $\beta_1 = 0.5$, and $\beta_2 = 0.9$. We train the CSA-Flow model for 240 epochs on MVTec-AD and BTAD datasets. For the real-world HSRBD datasets, we train the model for 480 epochs. The training process was performed using an NVIDIA RTX 3060 12G GPU.

Table 1: Comparison of area under ROC in % (AUROC) of different methods on MVTec-AD.

Method	STFPM	GANomaly	SPADE	PaDiM (R18-Rd100)	DifferNet	CS-Flow	CSA-Flow (ours)
Carpet	98.8	69.9	97.5	98.9	92.9	**100.0**	**100.0**
Grid	**99.0**	70.8	93.7	94.9	84.0	**99.0**	98.7
Leather	99.3	84.2	97.6	99.1	97.1	**100.0**	**100.0**
Tile	97.4	79.4	87.4	91.2	99.4	**100.0**	99.1
Wood	97.2	83.4	88.5	93.6	99.8	**100.0**	99.2
Texture classes	98.3	77.5	92.9	95.5	94.6	**99.8**	99.4
Bottle	98.8	89.2	98.4	98.1	99.0	**99.8**	**99.8**
Cable	95.5	75.7	97.2	95.8	95.9	**99.1**	98.7
Capsule	98.3	73.2	99.0	98.3	86.9	97.1	**99.4**
Hazelnut	98.5	78.5	99.1	97.7	99.3	99.6	**100.0**
Metal nut	97.6	70.0	98.1	96.7	96.1	**99.1**	98.1
Pill	97.8	74.3	96.5	94.7	88.8	**98.6**	97.5
Screw	98.3	74.6	**98.9**	97.4	96.3	97.6	96.9
Toothbrush	**98.9**	65.3	97.9	98.7	98.6	91.9	94.2
Transistor	82.5	79.2	94.1	97.2	91.1	99.3	**99.5**
Zipper	98.5	74.5	96.5	98.2	95.1	**99.7**	**99.7**
Object classes	96.5	75.5	97.6	97.3	94.7	98.2	**98.4**
Average	97.1	76.1	96.0	96.7	94.7	**98.7**	**98.7**

Bold values denote the best result in the category.

Table 2: Anomaly detection performance of AUROC (%) on the BTAD dataset.

Categories	AE MSE	AE MSE + SSIM	VT-ADL [20]	CSA-Flow (ours)
0	49	53	99	**99.61 (0.62%)**
1	92	**96**	94	87.20 (−9.17%)
2	95	89	77	**99.93 (5.19%)**
Average	78.67	79.33	90.00	**95.58 (6.20%)**

Note: We compared CSA-Flow with convolutional autoencoders trained with MSE-loss and MSE + SSIM loss. Bold values denote the best result in the category.

4.3. Anomaly Detection. We conducted experiments on the MVTec-AD datasets, which consist of 10 classes of objects and 5 classes of textures, to evaluate the performance of CSA-Flow. The training set exclusively contains defect-free images, while the test set includes both normal and abnormal images. We compared the performance of CSA-Flow with other anomaly detection models, including STFPM [39], GANomaly [25], SPADE [40], PaDiM [28], DifferNet [19], CS-Flow [32], using the AUROC metric.

The results, as shown in Table 1, demonstrate that the CSA-Flow model outperformed or achieved comparable performance to previous models in nearly half of the classes. Particularly, in terms of AUROC scores, CSA-Flow exhibited excellent performance compared to other reconstruction-based methods. In the research target of this paper, we should pay more attention to object classes, because it is more consistent with the goal of high-speed rail inspection.

In Table 2, we present a comparison between the basic convolutional AE using MSE and MSE + SSIM losses. The results demonstrate that CSA-Flow performs on par with MVTec-AD in terms of anomaly detection performance on the BTAD dataset.

In the HSRBD datasets, we conducted tests on four different scenarios of real-world HSRBD to evaluate the performance of the CSA-Flow model. Remarkably, the CSA-Flow model achieved the highest AUROC score compared to other models. To provide a more comprehensive evaluation, we proposed the use of Recall_Ano as a metric to assess the models' ability to detect abnormal samples. In industrial applications, detecting anomalies holds greater significance, considering the challenges posed by complex backgrounds and unclear subjects. Consequently, we believe it is essential to employ Recall as an evaluation metric.

We tested the AUROC and Recall_Ano on HSRBD datasets, and the results are shown in Tables 3 and 4. Notably, the CSA-Flow model outperforms previous models in the context of industrial applications. Figure 6 shows the accuracy comparison between CS-Flow and CSA-Flow in the HSRBD dataset. Our CSA-Flow model is significantly better than the original network.

These results suggest that existing methods struggle to effectively detect anomalies in scenarios with complex backgrounds and unclear subjects. In contrast, the proposed CSA-Flow model demonstrates outstanding performance in such industrial settings.

The primary goal of anomaly detection is not only to classify anomalies but also to segment abnormal parts. While the CSA-Flow model does not perform pixel-level evaluation, it uses anomaly scores to identify and locate defect regions. By analyzing these scores, we can effectively identify

Table 3: The AUROC score (%) on HSRBD datasets compared with the previous method.

Classes	GANomaly	SPADE	PaDiM	DifferNet	CS-Flow	CSA-Flow (ours)
Scenario 1	64.70	36.10	**89.20**	40.11	57.23	82.23 (−7.81%)
Scenario 2	68.70	86.00	96.00	69.56	98.67	**100.00 (1.35%)**
Scenario 3	76.00	77.40	96.60	73.56	94.23	**100.00 (3.52%)**
Scenario 4	57.00	**94.50**	67.30	60.61	83.03	89.09 (−5.72%)
Average	66.60	73.50	87.28	60.96	83.29	**92.83 (6.37%)**

Bold values denote the best result in the class.

Table 4: The Recall_Ano score (%) on the HSRBD datasets reflects the anomaly detection rate when compared to previous methods.

Classes	GANomaly	SPADE	PaDiM	DifferNet	CS-Flow	CSA-Flow (ours)
Scenario 1	65.40	3.85	80.77	80.77	61.54	**84.62 (4.76%)**
Scenario 2	60.00	60.00	**90.00**	**90.00**	**90.00**	**90.00 (0.00%)**
Scenario 3	87.50	75.00	**87.50**	75.00	**87.50**	**87.50 (0.00%)**
Scenario 4	73.30	**90.91**	72.73	18.18	63.64	81.82 (−10.00%)
Average	71.55	57.44	82.75	65.99	75.67	**85.98 (3.91%)**

Bold values denote the best result in the class.

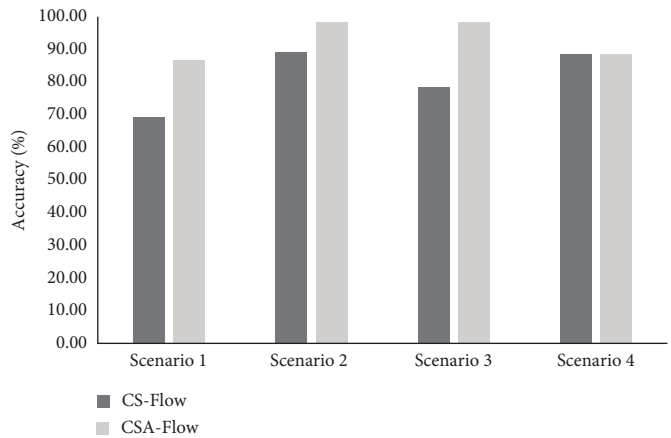

Figure 6: Accuracy comparison of CS-Flow and CSA-Flow in the HSRBD datasets.

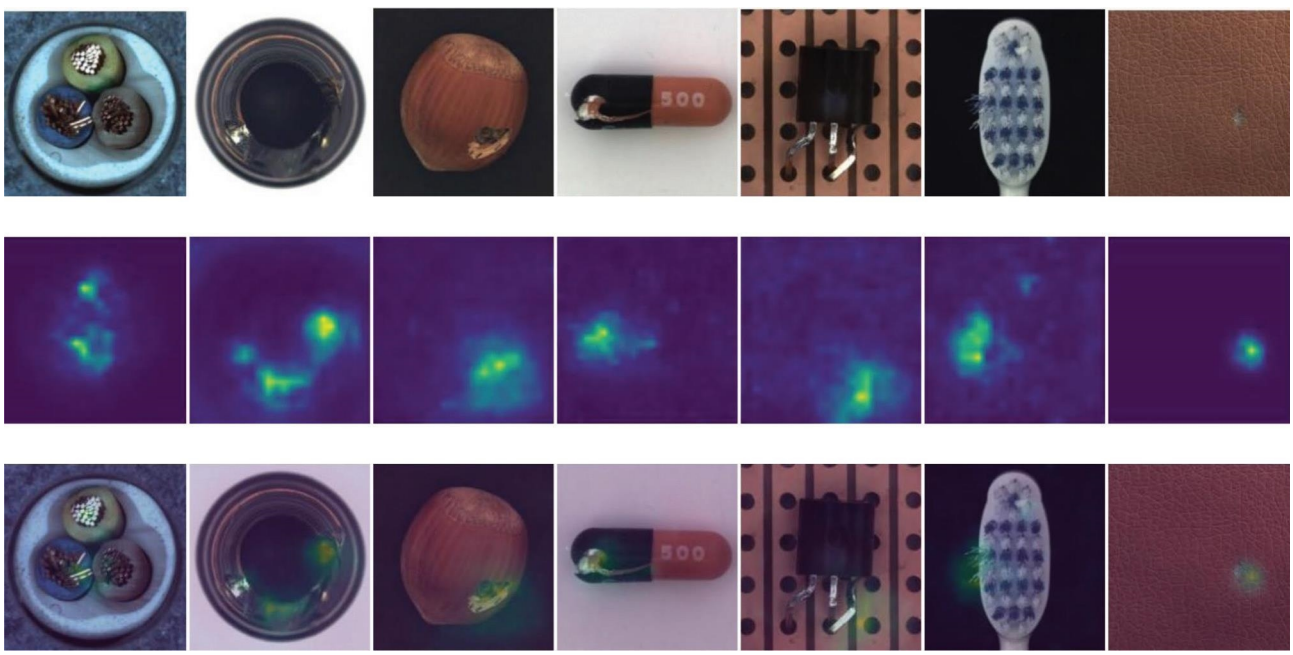

Figure 7: Examples of test results for various classes in the MVTec-AD dataset. The first row displays the original abnormal images, while the second row showcases the heatmaps of the defects. The bottom row provides visualizations of the predicted results for the defects.

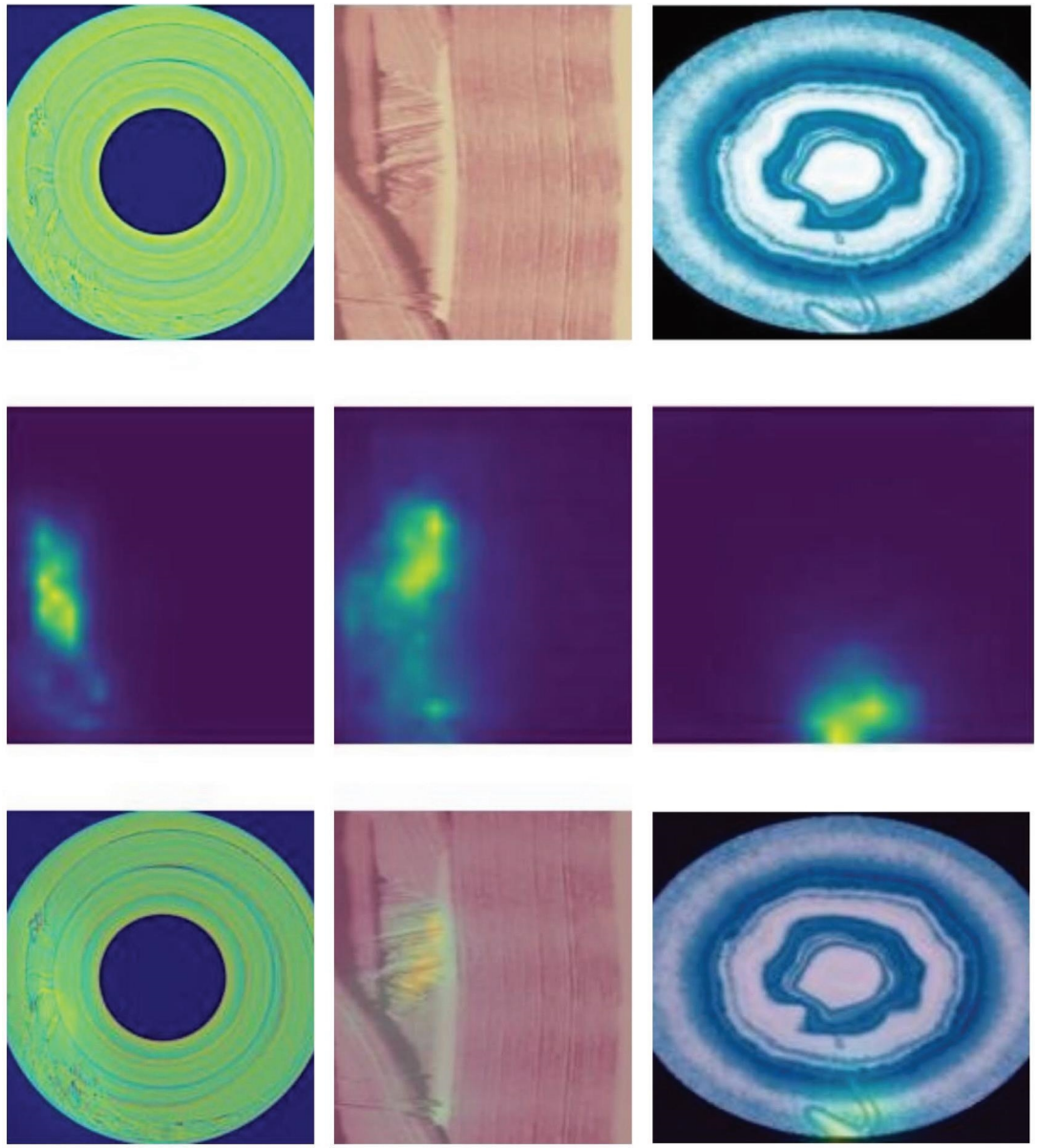

Figure 8: Examples of test results for various classes in the BTAD dataset.

abnormal areas. In the HSRBD datasets, where moving parts are considered normal, CSA-Flow demonstrates robustness and aligns with real-world scenarios. Although CSA-Flow is not explicitly designed for pixel segmentation, we assign anomaly scores to local positions (i, j) of the feature graph y^s by aggregating values along the channel dimension using $\left\| z^s_{i,j} \right\|^2_2$. By leveraging the high norms in the output tensorsz^s, we can accurately locate defects and assess them quickly. Figure 7 showcases localization in MVTec-AD, Figure 8 demonstrates localization in BTAD, and Figure 9 exhibits localization in HSRBD, highlighting CSA-Flow's accurate localization performance, particularly in industrial settings.

4.4. Ablation Study. To assess the effectiveness of the attentional module in our model, we conducted ablation experiments involving different subnetworks combinations. Specifically, we compared the AUROC metric, recall, and accuracy scores by including both the CA and SA modules and added only one of the modules for the HSRBD datasets. The results of these experiments are presented in Table 5. Since the HSRBD datasets are collected based on real train operation, foreign bodies with a diameter of less than 10 mm might cause 5.57% redundancy in the accuracy. Via comparison, it is worth noting that the inclusion of both CA and SA modules significantly enhances the accuracy of defect recognition in real-world industrial defect scenarios to improve the detection efficiency of high-speed electric multiple unit (EMU).

In industrial detection, anomaly detection is widely applied to the primary maintenance of high-speed EMU. The time of this primary maintenance is urgent, and the current process regulations are constantly reducing the

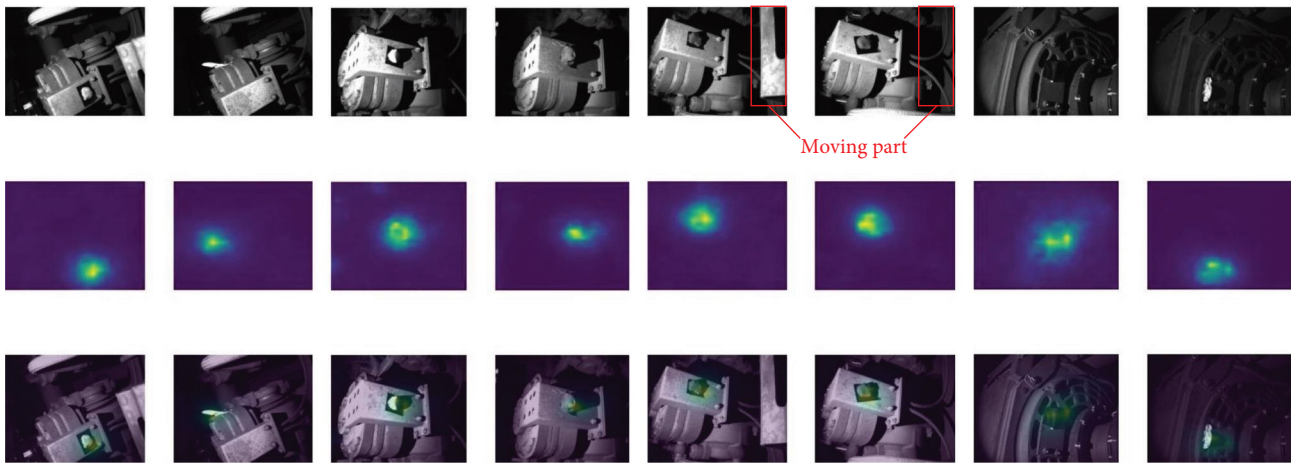

Figure 9: Examples of test results for various classes in the HSRBD datasets. The results demonstrate the excellent segmentation performance for detecting foreign bodies and distinguishing moving parts, which should be considered normal in this context.

Table 5: Ablation was studied on HSRBD datasets on average by adding different modules.

	AUROC (%)	Recall_Norm (%)	Recall_Ano (%)	Accuracy (%)
CSFlow	83.4	77.1	79.7	77.6
CA	87.60	87.06	78.08	83.85
SA	80.07	74.58	79.68	75.12
CSA-Flow	**94.08**	**98.41**	**83.53**	**94.43**

Bold values denote the best result by using which module.

maintenance time; therefore, the workload of manual review can be greatly reduced due to the improvement of accuracy, especially if the base is large.

5. Conclusion

In this paper, we presented CSA-Flow, which combines cross-scale NF with attention modules and applied to a practical application of high-speed EMU. We aim to improve anomaly detection accuracy and reduce the workload of manual review. We also introduced the channel feature extraction module for different scales of feature extraction, and our experiments demonstrate the promising potential of CSA-Flow. We believe that evaluating the performance of a network in industrial applications is crucial. Existing networks often face challenges in industrial settings due to factors such as lighting, background information, and texture, which can impact detection results. As shown in Table 2, our proposed method excels at detecting foreign bodies in complex backgrounds. We also introduced more intuitive metrics that are highly relevant in industrial applications, such as the detection rate. Consequently, we evaluated the recall of CSA-Flow on the HSRBD datasets, and the results demonstrated that our method achieved the highest anomaly detection rate. Although CSA-Flow does not perform precise pixel-level segmentation, we can utilize the anomaly scores to locate abnormal parts. Future research will focus on improving speed and advancing pixel-level segmentation capabilities.

Data Availability

Data underlying the result presented in this paper are available in [14, 20]. Other generated data are not publicly available at this time but may be obtained from the authors upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (grant no. 61960206010).

References

[1] G. Piao, J. Li, L. Udpa, J. Qian, and Y. Deng, "Finite-element study of motion-induced eddy current array method for high-speed rail defects detection," *IEEE Transactions on Magnetics*, vol. 57, no. 12, pp. 1–10, 2021.

[2] D. F. Cannon, K.-O. Edel, S. L. Grassie, and K. Sawley, "Rail defects: an overview," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 26, no. 10, pp. 865–886, 2003.

[3] S. Liu, Q. Wang, and Y. Luo, "A review of applications of visual inspection technology based on image processing in the railway industry," *Transportation Safety and Environment*, vol. 1, no. 3, pp. 185–204, 2019.

[4] Z. Teng, F. Liu, and B. Zhang, "Visual railway detection by superpixel based intracellular decisions," *Multimedia Tools and Applications*, vol. 75, no. 5, pp. 2473–2486, 2016.

[5] T. Czimmermann, G. Ciuti, M. Milazzo et al., "Visual-based defect detection and classification approaches for industrial applications—a survey," *Sensors*, vol. 20, no. 5, Article ID 1459, 2020.

[6] P. L. Mazzeo, M. Nitti, E. Stella, and A. Distante, "Visual recognition of fastening bolts for railroad maintenance," *Pattern Recognition Letters*, vol. 25, no. 6, pp. 669–677, 2004.

[7] F. Marino, A. Distante, P. L. Mazzeo, and E. Stella, "A real-time visual inspection system for railway maintenance: automatic hexagonal-headed bolts detection," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 418–428, 2007.

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

[9] J. Sun, Z. Xiao, and Y. Xie, "Automatic multi-fault recognition in TFDS based on convolutional neural network," *Neurocomputing*, vol. 222, pp. 127–136, 2017.

[10] A. Ma, Z. Lv, X. Chen et al., "Pandrol track fastener defect detection based on local convolutional neural networks," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 235, no. 10, pp. 1906–1915, 2021.

[11] Y. Xu, F. Sun, and L. Wang, "YOLOv5-PD: a model for common asphalt pavement defects detection," *Journal of Sensors*, vol. 2022, Article ID 7530361, 12 pages, 2022.

[12] M. Yang, P. Wu, and H. Feng, "MemSeg: a semi-supervised method for image surface defect detection using differences and commonalities," *Engineering Applications of Artificial Intelligence*, vol. 119, Article ID 105835, 2023.

[13] Y. He, J. Wu, Y. Zheng, Y. Zhang, and X. Hong, "Track defect detection for high-speed maglev trains via deep learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–8, 2022.

[14] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

[15] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[16] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, "Normalizing flows: an introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, 2021.

[17] M. A. Cho, T. Kim, W. J. Kim, S. Cho, and S. Lee, "Unsupervised video anomaly detection via normalizing flows with implicit latent features," *Pattern Recognition*, vol. 129, Article ID 108703, 2022.

[18] L. Dinh, D. Krueger, and Y. Bengio, "Nice: non-linear independent components estimation," arXiv preprint arXiv: 1410.8516, 2014.

[19] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but DifferNet: semi-supervised defect detection with normalizing flows," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1907–1916, 2021.

[20] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "VT-ADL: a vision transformer network for image anomaly detection and localization," in *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 1–6, IEEE, June 2021.

[21] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25–30, 2017, Proceedings*, pp. 146–157, Springer, Cham, May 2017.

[22] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Artificial Neural Networks and Machine Learning-ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14–17, 2011, Proceedings, Part I 21*, pp. 44–51, Springer, Berlin, 2011.

[23] M. Rudolph, B. Wandt, and B. Rosenhahn, "Structuring autoencoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[24] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient gan-based anomaly detection," arXiv preprint arXiv: 1802.06222, 2018.

[25] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: semi-supervised anomaly detection via adversarial training," in *Computer Vision-ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, vol. III 14, pp. 622–637, Springer, Perth, Australia, 2019.

[26] L. Bergman and Y. Hoshen, "Classification-based anomaly detection for general data," arXiv preprint arXiv: 2005.02359, 2020.

[27] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4183–4192, 2020.

[28] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 2021, Proceedings, Part IV*, pp. 475–489, Springer, Cham, March 2021.

[29] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: an introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3964–3979, 2021.

[30] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," arXiv preprint arXiv: 1605.08803, 2016.

[31] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 98–107, 2022.

[32] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1088–1097, 2022.

[33] L. Bergman, N. Cohen, and Y. Hoshen, "Deep nearest neighbor anomaly detection," arXiv preprint arXiv: 2002.10445, 2020.

[34] A. A. Bastidas and H. Tang, "Channel attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[35] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: frequency channel attention networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 783–792, 2021.

[36] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[37] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[38] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv preprint arXiv: 1412.6980, 2014.
[39] G. Wang, S. Han, E. Ding, and D. Huang, "Student–teacher feature pyramid matching for anomaly detection," arXiv preprint arXiv: 2103.04257, 2021.
[40] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," arXiv preprint arXiv: 2005.02357, 2020.