Research Article

AnoDFDNet: A Deep Feature Difference Network for Anomaly Detection

Zhixue Wang*, Yu Zhang*, Lin Luo*, and Nan Wang*

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

Correspondence should be addressed to Yu Zhang; zhang.yuer@163.com

Received 8 June 2022; Accepted 1 August 2022; Published 16 August 2022

Academic Editor: Sangsoon Lim

Copyright © 2022 Zhixue Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper proposed a novel anomaly detection (AD) approach of high-speed train images based on convolutional neural networks and the Vision Transformer. Different from previous AD works, in which anomalies are identified with a single image using classification, segmentation, or object detection methods, the proposed method detects abnormal difference between two images taken at different times of the same region. In other words, we cast anomaly detection problem with a single image into a difference detection problem with two images. The core idea of the proposed method is that the “anomaly” commonly represents an abnormal state instead of a specific object, and this state should be identified by a pair of images. In addition, we introduced a deep feature difference AD network (AnoDFDNet) which sufficiently explored the potential of the Vision Transformer and convolutional neural networks. To verify the effectiveness of the proposed AnoDFDNet, we gathered three datasets, a difference dataset (Diff dataset), a foreign body dataset (FB dataset), and an oil leakage dataset (OL dataset). Experimental results on the above datasets demonstrate the superiority of the proposed method. In terms of the F1-score, the AnoDFDNet obtained 76.24%, 81.04%, and 83.92% on Diff dataset, FB dataset, and OL dataset, respectively.

1. Introduction

Anomaly detection (AD) is one of the core tasks in computer vision which has been well-studied within a wide range of research areas and application domains. Its critical idea is how to identify abnormal information that significantly deviates from the majority of data information [1, 2]. Depending on different application situation or data types, an anomaly is also known as outlier or novelty [3]. Approaches and applications of anomaly detection exist in various domains, such as fraud detection [4], video surveillance [5, 6], health-care [7], security check [8], fault detection [9, 10], and defect detection [11–20].

In the field of safety inspection of high-speed trains, AD is widely used to identify defects and anomalies. Due to the extreme importance of train safety inspection in ensuring the safe and reliable operation of high-speed trains and the development of machine vision, more and more applications based on AD task are used in train safety inspection to improve detection efficiency, reduce detecting cost, and realize intelligent detection [13–17]. In these methods, most of them focus on how to detect anomalies and identify surface defects of key components. Currently, inspired by the success of convolutional neural networks (CNNs) on image analysis, more and more AD algorithms are equipped with CNNs to meet the requirement of fast computing speed, high efficiency, and detecting accuracy [18–27]. According to the image analysis types, the popular AD methods for high-speed train safety inspection can be divided into four categories, unsupervised generated methods [7, 8, 11, 12], anomaly or defect classification [17, 22, 25, 28–30], abnormal object detection [9, 10, 18–21, 23, 26], and defect segmentation [29, 31–33].

Even though the above methods have achieved considerable performance in detecting anomalies, we believe that using object-based methods to detect anomalies is not a very appropriate way. The main reason lies in three respects:

(1) The “anomaly” generally represents an abnormal state instead of a specific object. As shown in Figure 1, we illustrated some abnormal samples taken from the bottom of high-speed trains. For
Figure 1: Continued.
samples (a), (d), (f), and (c), it is available that utilizing an object detection or segmentation model to detect “foreign body” and “scratch.” In this case, the region of interest (ROI) remains a specific abnormal object to be detected. Nonetheless, for samples (b) and (e), since the anomaly is caused by a missed component instead of an added abnormal object, object detection or segmentation method does not work on it. In other words, in this region, there are no extraneous abnormal objects that can be localized or identified. Consequently, the object-based detection methods are ineffective to identify anomalies if there are no abnormal objects.

(2) Each object usually has obvious features that represent itself as belonging to a category. In AD tasks, different abnormal categories have different abnormal objects, and some of the objects in each abnormal category have very different feature representation. As samples (a), (d), and (f) shown in Figure 1, the category “foreign body” has different object types, such as tissue, packaging tape, or other unlisted abnormal objects. These abnormal objects have different sizes, shapes, and less structure features. This would make it difficult for a pretrained model to detect unseen abnormal objects whose geometric appearance is very different from training data. In addition, one thing we cannot ignore is that, due to the lack of abnormal samples, it is very difficult to train a powerful enough model that is able to detect all the potential abnormal objects.

(3) During the train operation, it might have other foreign bodies, such as cables, tree branches, the bodies of animals or birds, stones, plastic bags, mud, or ice and snow which may threaten train operation safety, and so on. It is impossible to list all anomaly categories and abnormal object types. Therefore, there is no doubt that a pretrained model will fail to detect unseen anomalies and abnormal objects. For example, a “foreign body” dataset contains samples of stones, plastic bags, and tissues. The pretrained model will have the capacity of detecting these abnormal objects. For the unseen foreign bodies such as tree branches, animal, or bird bodies, the model did not see them in the training phase and they have different representation features from trained objects, they would be misdetected.

In this paper, we proposed a novel AD method. Unlike previous methods, we use two images taken at different times to detect anomalies by differentiating their features. To the best of our knowledge, the use of paired images and Siamese architectures for AD task is an area of research that has only been poorly studied. As shown in Figure 2, a pair of samples contain two images taken at different times in the same region of the same train. For convenience, the image acquired at the previous time is called as the “history image,”
and the image acquired at the latest time is denoted as the “current image.” As we mentioned above, the “anomaly” describes a state instead of a specific object, and some anomalies can not be identified only from a single image only. Therefore, it is much easier to identify whether the ROI exists “anomaly” using two images’ comparison than only using a single image [25]. No matter what type of the abnormal object it is, there must remain a difference between the normal and abnormal images. This difference indicates the abnormal information. At this point, the AD task turns into a problem of detecting whether a pair of samples have interesting differences. Note that a preprint
of the proposed method has previously been published [34]. Source code is available at https://github.com/wangle53/AnoDFDNet.

The overview of the proposed model can be seen from Figure 2. This model consists of four components: two CNNs are used to extract the input image features; a Transformer-based module is used to establish long-range relationships; a deep feature difference decoder is used to generate anomaly maps.

The contribution can be summarized as follows:

1. This paper proposed a novel AD method that accepts two images as input to detect anomalies, which casts abnormal object detection problem to a difference detection problem. Transforming the problem of inexhaustibility of anomalies into a dichotomous problem of whether there exists a difference

2. A deep feature difference AD model named AnoDFDNet is designed to deliver AD task, which

<table>
<thead>
<tr>
<th>Datasets</th>
<th>No.</th>
<th>Methods</th>
<th>Pre</th>
<th>Re</th>
<th>OA</th>
<th>F1</th>
<th>IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff dataset</td>
<td>1</td>
<td>Unet [38]</td>
<td>0.4867</td>
<td>0.1021</td>
<td>0.9878</td>
<td>0.1688</td>
<td>0.0922</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Unet-cat [38]</td>
<td>0.3107</td>
<td>0.2707</td>
<td>0.9838</td>
<td>0.2893</td>
<td>0.1691</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>DeeplabV3+ [39]</td>
<td>0.3834</td>
<td>0.2674</td>
<td>0.9863</td>
<td>0.3151</td>
<td>0.1870</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>DeeplabV3+-cat [39]</td>
<td>0.3075</td>
<td>0.6135</td>
<td>0.9785</td>
<td>0.4097</td>
<td>0.2576</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>FCLNet [40]</td>
<td>0.4938</td>
<td>0.5075</td>
<td>0.9881</td>
<td>0.5006</td>
<td>0.3338</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>AnoDFDNet</td>
<td>0.7416</td>
<td>0.7584</td>
<td>0.9940</td>
<td>0.7499</td>
<td>0.5999</td>
</tr>
<tr>
<td>FB dataset</td>
<td>7</td>
<td>Unet [38]</td>
<td>0.5667</td>
<td>0.5617</td>
<td>0.8572</td>
<td>0.5642</td>
<td>0.3930</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Unet-cat [38]</td>
<td>0.7975</td>
<td>0.6452</td>
<td>0.9146</td>
<td>0.7133</td>
<td>0.5544</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>DeeplabV3+ [39]</td>
<td>0.7624</td>
<td>0.5173</td>
<td>0.8940</td>
<td>0.6164</td>
<td>0.4455</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>DeeplabV3+-cat [39]</td>
<td>0.6855</td>
<td>0.8604</td>
<td>0.9120</td>
<td>0.7630</td>
<td>0.6169</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>FCLNet [40]</td>
<td>0.7777</td>
<td>0.7151</td>
<td>0.9206</td>
<td>0.7451</td>
<td>0.5937</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>AnoDFDNet</td>
<td>0.8328</td>
<td>0.7891</td>
<td>0.9392</td>
<td>0.8104</td>
<td>0.6812</td>
</tr>
<tr>
<td>OL dataset</td>
<td>13</td>
<td>Unet [38]</td>
<td>0.8118</td>
<td>0.7589</td>
<td>0.9518</td>
<td>0.7845</td>
<td>0.6454</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Unet-cat [38]</td>
<td>0.8022</td>
<td>0.6111</td>
<td>0.9376</td>
<td>0.6937</td>
<td>0.5310</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>DeeplabV3+ [39]</td>
<td>0.7674</td>
<td>0.8361</td>
<td>0.9507</td>
<td>0.8003</td>
<td>0.6671</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>DeeplabV3+-cat [39]</td>
<td>0.8086</td>
<td>0.7921</td>
<td>0.9533</td>
<td>0.8003</td>
<td>0.6671</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>FCLNet [40]</td>
<td>0.7951</td>
<td>0.8250</td>
<td>0.9563</td>
<td>0.8098</td>
<td>0.6804</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>AnoDFDNet</td>
<td>0.8395</td>
<td>0.8227</td>
<td>0.9613</td>
<td>0.8310</td>
<td>0.7109</td>
</tr>
</tbody>
</table>

Figure 4: Typical samples with large viewpoint difference. From (a) to (d) in each row, they are the "cur" image, the "his" image, the ground truth, and the fusion of both images, respectively.
sufficiently explored the potential of Vision Transformer and CNNs.

(3) We collected three AD datasets of high-speed trains to evaluate the effectiveness of the proposed AnoDFDNet, a difference detection dataset (Diff dataset), a foreign body dataset (FB dataset), and an oil leakage dataset (OL dataset). The proposed AnoDFDNet achieved 76.24%, 81.04%, and 83.92% in terms of the F1-score, respectively.

The rest of this paper is organized as follows: Section 2 explores other works that were used for inspiration or comparison during the development of this work. The proposed method is described in Section 3. The datasets and experiment configuration are described in Section 4. Section 5 shows the experimental results. Finally, the conclusion of this paper is drawn in Section 6.

2. Related Works

Recently, inspired by the high performance achieved by CNNs in image tasks, many CNN-based methods have been introduced in the AD task. In view of the lack of abnormal data, the unsupervised methods firstly train a model only on samples considered to be normal and then identify
Table 2: Comparison of inference speed (frame per second (FPS)).

<table>
<thead>
<tr>
<th>No.</th>
<th>Methods</th>
<th>Layer number</th>
<th>FPS (CPU)</th>
<th>FPS (GPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unet [38]</td>
<td>—</td>
<td>8.61</td>
<td>102.14</td>
</tr>
<tr>
<td>2</td>
<td>DeeplabV3+ [39]</td>
<td>—</td>
<td>3.65</td>
<td>48.47</td>
</tr>
<tr>
<td>3</td>
<td>FCLN [40]</td>
<td>—</td>
<td>1.15</td>
<td>50.49</td>
</tr>
<tr>
<td>4</td>
<td>AnoDFDNet</td>
<td>0</td>
<td>6.86</td>
<td>124.60</td>
</tr>
<tr>
<td>5</td>
<td>AnoDFDNet</td>
<td>1</td>
<td>5.51</td>
<td>95.81</td>
</tr>
<tr>
<td>6</td>
<td>AnoDFDNet</td>
<td>2</td>
<td>4.62</td>
<td>88.36</td>
</tr>
<tr>
<td>7</td>
<td>AnoDFDNet</td>
<td>4</td>
<td>3.48</td>
<td>78.35</td>
</tr>
<tr>
<td>8</td>
<td>AnoDFDNet</td>
<td>8</td>
<td>2.40</td>
<td>63.62</td>
</tr>
</tbody>
</table>

insufficiently abnormal samples which differ from the learned data distribution of normal data [7, 8]. Salehi et al. [11] used knowledge distillation to learn a cloner network from an expert network. This method detects and localizes anomalies using the discrepancy between the expert and cloner networks’ intermediate activation values. Bergmann et al. [12] proposed a student-teacher anomaly detection model. It detects anomalies according to the output difference between the student network and the teacher network. For anomaly classification, the task comes to directly classifying different types of anomalies or defects. Wang et al. [25] proposed three kinds of classification models to classify paired images. In their dataset, different anomalies are divided into six categories. As the same as [25], Giben et al. [29] represented a classification method to identify different material of railway track images. For abnormal object detection, it usually takes two steps: first localizing the key component or the region needing to be detected and second using a classifier to identify whether this region or key component is abnormal or which category it belongs to. Chen et al. [18] proposed a coarse-to-fine detecting method. Their networks consist of three subnetworks, two detectors sequentially localize the cantilever joints and their fasteners, and a classifier to diagnose the fasteners’ defects. According to different defect types, the fasteners are classified into three categories, “normal,” “missing,” and “latent missing”. Tu et al. [23] proposed a real-time defect detection method of track components. In this method, they take the influence of sample imbalance and subtle difference among different classes into consideration. Segmentation-based methods of AD task are usually used to segment defects. Further process is using a classifier to classify the segmented defects into different types. Cao et al. [32] proposed a pixel-level segmentation network for surface defect detection. To improve the model performance, the authors integrated feature aggregation and attention modules into the network. Their experiments demonstrated that this model achieved good performance. Li et al. [33] presented a semantic-segmentation-based algorithm for the state recognition of rail fasteners. In addition, the pyramid scene analysis network and vector geometry measurements are combined into the network to further improve its performance.

From the above works, it can be observed that the most existing approaches often focus on detecting specific anomalies. Therefore, these works cannot be applied to unseen anomalies which did not be studied by the network. As mentioned in Section 1, this is the motivation why we designed the proposed method. Since all anomalies can be identified by two image comparison, using a feature difference network can overcome above drawbacks and regardless of the specific categories of anomalies.

3. Proposed Method

The overview architecture of the proposed AnoDFDNet is presented in Figure 2. This model consists of four parts: two weight-shared CNNs used to extract image features; a Vision Transformer-based [35–37] module used to establish long-range relationships and reform feature representation; the deep feature difference decoder used to fuse differentiated features and generate final detecting results. The corresponding layers between CNNs and the decoder are connected using skip connection.

Given two images $\text{cur}, h_i \in \mathbb{R}^{H \times W \times C}$, acquired at different times of the same region of the same train, where $H \times W$ denotes its spatial resolution and $C$ denotes the channel dimension. First, input images are fed into CNNs to extract feature representation:

$$F_c = \text{CNN}(\text{cur}),$$
$$F_h = \text{CNN}(\text{his}),$$
$$F_c, F_h \in \mathbb{R}^{H \times W \times C}.$$  

Before feeding feature representation into Transformer, we calculate their feature difference using absolute subtraction operation.

$$F = \text{abs}(F_c - F_h).$$

The differentiated features $F \in \mathbb{R}^{H \times W \times C}$ are converted into token sequences $p_i \in \mathbb{R}^d$ using transpose, reshape operation. Then, we map them into a latent $d$-dimensional embedding space using a trainable linear projection. After that, a learnable position embedding $E_{\text{pos}} \in \mathbb{R}^{d \times \text{len}}$, which is able to retain positional information is added to token sequences.

$$z_0 = [E(p_1) ; E(p_2) ; E(p_3) ; \cdots ; E(p_L)] + E_{\text{pos}},$$

where $l = h \times w$ presents the length of token sequences and $E(p_i) \in \mathbb{R}^d$ denotes the $i$th token sequence after linear projection.

The proposed AnoDFDNet consists of $L$ layers of Transformer. As shown in Figure 2, each Transformer is stocked with a multihead self-attention (MSA) block, a multilayer perceptron (MLP) block, and two layer-normalization (LN) layers. At each layer $\ell$, its output can be described as follows:

$$z_\ell' = \text{MSA} (\text{LN}(z_{\ell-1})), \quad \ell = 1 \cdots L,$$
$$z_\ell = \text{MLP} \left( \text{LN}(z_\ell') \right) + z_\ell', \quad \ell = 1 \cdots L.$$  

The decoder followed by Transformer layers is used to
<table>
<thead>
<tr>
<th></th>
<th>Cur</th>
<th>His</th>
<th>Ground truth</th>
<th>Unet</th>
<th>Unet_cat</th>
<th>Deeplabv3+</th>
<th>Deeplabv3+_cat</th>
<th>FCLNet</th>
<th>AnoDFDNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cur dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FB dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>OL dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6: Visualization of the anomaly maps of different methods on three datasets. From (a) to (i) in each row, the image denotes "cur" image, "his" images, ground truth, and the anomaly map obtained by Unet, Unet-cat, Deeplabv3+, Deeplabv3+_cat, FCLNet, and our AnoDFDNet, respectively. Here, the white pixel denotes detected anomalies, and the black denotes the background without anomalies.
be formulated as follows:

\[
\text{Loss} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} -y_{i,j} \log a_{i,j} + (1 - y_{i,j}) \times \log (1 - a_{i,j}).
\]  

(6)

4. Experimental Implementation Details

4.1. Datasets. To evaluate the performance of the proposed AnoDFDNet, we collected three challenging anomaly detection datasets, a difference detection dataset (Diff dataset), a foreign body detection dataset (FB dataset), and an oil leakage dataset (OL dataset).

4.1.1. Diff Dataset. As shown in Figure 1, this dataset contains 399 pairs of training and validation samples and 101 pairs of testing samples. All images are captured automatically by a robot with a camera. By walking under the high-speed train, it can take images of specific regions and components. The main anomalies of this dataset are foreign bodies, missing components, loose components, detached components, scratches, and other anomalies. The spatial resolutions of each image are all 1920 × 1200. Training dataset is split into training and validation sets with the ratio of 8:2. All images are scaled to 256 × 256 before being fed into network.

4.1.2. FB Dataset. This dataset is a collection of foreign body anomalies and contains 180 pairs of training and validation samples and 39 pairs of testing samples; some of the samples are shown in Figure 3(a). Their spatial resolutions vary from 123 × 271 to 4247 × 1282. To be the same with Diff dataset’s operation, the dataset is split into training and validation with the ratio of 8:2, and all images are scaled to 256 × 256.

4.1.3. OL Dataset. This dataset is a collection of oil leakage anomalies, and some samples are shown in Figure 3(b). It consists of 1275 pairs of training and validation samples and 153 pairs of testing samples with the same resolution of 2048 × 1536. Follow the same operation of above two datasets, the dataset is split into training and validation with the ratio of 8:2, and all images are scaled to 256 × 256.

It is worth noting that all above datasets have many noisy differences, such as illumination variation, component rotation, and viewpoint variation. These noisy differences will have a certain impact on the anomaly detection. For an AD model, the critical idea lies in if it can detect true anomaly difference while rejecting noisy differences.

4.2. Optimization and Evaluation. In our experiments, all networks were trained using the Adam algorithm with a learning rate of 2e-4. All experiments were implemented using PyTorch 1.10.0 and with an Nvidia RTX2070 GPU of 8 G memory and Intel i7-8700 CPU. To evaluate the performance of the proposed AnoDFDNet, five evaluation measures are used, the precision (Pre), recall (Re), overall accuracy (OA), F1-score (F1), and intersection over union (IoU).

5. Experiments

5.1. Comparison Experiments. In this section, we compared the proposed model with several popular methods which
can be used to detect anomalies of paired images. To the best of our knowledge, it does not receive much attention on detecting anomalies using paired images. Therefore, we compared AnoDFDNet with two popular segmentation methods, Unet [38] and DeeplabV3+ [39]. In our view, the purpose of AD task is to generate an anomaly map; it is similar to image segmentation task except the inputs are paired images. Note that the inputs are two images, the paired images are concatenated into a 6-channel image (in Table 1, “_cat” denotes that two images are concatenated into a 6-channel image as input) before being fed into segmentation networks. In addition, since change detection is a task that can also process bi-temporal images, we selected FCLNet [40] as a comparison. In the following, if no otherwise specified, the claimed AnoDFDNet denotes the one with 2 Transformer layers.

As shown in Table 1, comparison experiments were demonstrated on three datasets. It can be observed that the proposed AnoDFDNet achieved superior performance over all datasets. On Diff dataset, the proposed model achieved
74.99% and 59.99% in terms of the F1-score and IoU, respectively. Compared with FCL, AnoDFDNet has improvements of 24.93% and 26.61% on the F1-score and IoU. On FB dataset, AnoDFDNet obtained 81.04% and 68.12% in terms of the F1-score and IoU and improved the performance by 6.53% and 8.75%. For OL dataset, the proposed model still achieved the superior performance with 83.10% in F1-score and 71.09% in IoU, which improved by 2.12% and 3.05%.

On both of Diff and FB datasets, the performances of Unet_cat and Deeplabv3+_cat are better than Unet and Deeplabv3+. There are three main reasons for this; one is two images provide more information for feature extraction; more importantly, the difference information only can be detected using two images acquired at different times; the third one is that anomaly objects are inexhaustible and cannot be listed in their entirety; for some specific anomalies, such as foreign bodies, they do not have a particular geometrical characteristic. Therefore, it is difficult to detect anomalies using a single image.

Another important point here is that the overall performance of OL dataset is higher than Diff and FB datasets. The main reason is that Diff and FB datasets have larger viewpoint than OL dataset. As shown in Figure 4, we fused “cur” and “his” image into an image. As the existence of viewpoint difference, two images are unable to match with each other. In the other word, the two images are unregistered. This unregistered viewpoint difference makes detecting anomalies be more difficult. From Table 1, on Diff dataset, the proposed AnoDFDNet improved with a huge margin compared with FCLNet. It indicates that AnoDFDNet is good to overcome viewpoint difference.

For further comparison, we figured out their receiver operating characteristic (ROC) curves. As shown in Figure 5, the proposed model obtained the highest AUC (area under curve) and the lowest EER (equal error rate). AnoDFDNet achieved 98.61%, 96.80%, and 98.73% in terms of AUC and 5.05%, 9.72%, and 5.46% in terms of EER, on Diff, FB, and OL datasets, respectively. Table 2 shows the inference speeds of the methods; we can observe that the proposed method has a good trade-off between detecting accuracy and computing complexity. From above comparison, it indicates that the proposed model has better capacity of detecting anomalies than other methods.

5.2. Visualization. In this section, we have visualized anomaly maps of the selected methods to visually compare them. As shown in Figure 6, on Diff dataset, it is no doubt that the presented AnoDFDNet performed the best. As the existence of large viewpoint difference between two images, all selected methods seem incompetent to detect anomalies. For example, for the fourth pair of samples, all selected methods missed the missing anomaly and the scratch.
anomaly; for the third pair of samples, Unet_cat detected the unregistered pipe fitting as anomalies instead of the foreign body. As to FB dataset, although its viewpoint difference is smaller than Diff dataset, it still brings difficulties to anomaly detection. For the second pair of samples, both Unet and Deeplabv3+ missed the foreign body anomaly; Unet_cat and Deeplabv3+_cat just detected a little part of the anomaly.

With regard to OL dataset, even though the performances of all methods are comparable, some difference can be found on closer observation. Unet missed the anomalies on the top left corner of the first pair of samples. For the third pair of samples, the anomaly map detected by FCLNet is not fine enough. In the last pair of samples, only our model did a good job of detecting the oil leakage anomaly in the box.

5.3. Ablation Studies. In the proposed AnoDFDNet, a significant module is the Transformer-based architecture. To investigate the influence of Transformer layers on model performance, we implemented a series of ablation experiments around the number of Transformer layers; here, “0” denotes without Transformer module in AnoDFDNet.

As shown in Table 3, on OL dataset, the results with different Transformer layers are similar; on FB dataset, the model with 2 layers of Transformer obtained the highest performance; on Diff dataset, the performance is superior to others when layer number is set to 8. More importantly, the results with or without Transformer are very different. For example, in terms of the IoU, experiment No. 5 improved by 10.28% when compared with No. 1; experiment No. 8 has an improvement of 2.49% over No 6; experiment No. 12 is 3.06% higher than No. 11. Form above results, it denotes that the designed Transformer-based module indeed improves the detecting performance. To some extent, the performance improvement benefits from the Transformer-based module especially on datasets with large viewpoint difference, because the Transformer is good at establishing global semantic relations and modeling long-range context which benefits to samples with large viewpoint difference. It is worth noting that, although experiment No. 7, No. 8, and No. 9 have superior performances than No. 6, experiment No. 10 has lower performance than No. 6 on FB dataset. One reasonable explanation is that the samples of FB dataset are not very enough to train a complicated Transformer-based model with 8 layers of Transformers.

The anomaly maps generated by AnoDFDNet without Transformer and with two layers of Transformers are listed in Figure 7. From Figure 7, it shows that Transformer-based AnoDFDNet outperforms the one without Transformer. On Diff dataset, for the first and second pairs of samples, no-Transformer AnoDFDNet identified part of background into anomalies; for the third pair of samples, no-Transformer AnoDFDNet missed some anomalies. On FB dataset, different from no-Transformer AnoDFDNet which detected much false positive, the Transformer-based AnoDFDNet obtained good anomaly maps on three pairs of samples. On OL dataset, as to the first pair of samples, no-Transformer AnoDFDNet missed a large abnormal region. From the above experiment results, it verifies the superiority of the resigned AnoDFDNet.

5.4. Visualization of Feature Maps and Dimensionality. To interpret the proposed AnoDFDNet as a deeper level, we visualized the feature maps extracted from the “d2” layer (as shown in Figure 2). In addition, the T-SNE (T-distributed Stochastic Neighbor Embedding) [41] algorithm is used to embed the feature distribution.

As shown in Figure 8, we selected 3 typical feature maps for each pair of samples. The feature map 1 mainly focuses on the background and the abnormal region’s boundary. Feature maps 2 and 3 response for the anomaly itself. All of them allow the model focus training on more useful feature representations. The last column visualized the three-dimensional feature embedding; it shows that the proposed model obtained a good distinction between abnormal and normal feature information.

6. Conclusions

In this paper, we proposed a novel anomaly detection framework named AnoDFDNet based on the fusion of convolutional neural networks and the Vision Transformer. In view of the drawbacks of detecting anomalies with a single image, we introduce the idea that the “anomaly” describes an abnormal state instead of a specific object. Therefore, it is more reasonable that the anomaly detection should be operated upon a pair of images. By detecting the interesting image differences, the abnormal information can be identified. Following above idea, we established a deep feature difference anomaly detection model, in which convolutional neural networks and the Vision Transformer are fused together. Extensive experiments indicate that the proposed method outperforms other comparison methods. In addition, the integrated Transformer layers indeed improve detection performance and benefit to detecting such samples with large viewpoint difference. In terms of the F1-score, the AnoDFDNet obtained 76.24%, 81.04%, and 83.92% on Diff dataset, FB dataset, and OL dataset, respectively. This paper provided a novel method of both researchers and practitioners to further develop anomaly detection task.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

A preprint version can be found from the link: https://arxiv.org/abs/2203.15195.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61771409 and the Science and Technology Program of Sichuan under Grant No. 2021YJ0080, for providing the project of research.
on high-speed train safety inspection. We gratefully acknowledge the arXiv for providing the opportunity of uploading our manuscript for peer communication.

References


