Article

A Lightweight Model for Bearing Fault Diagnosis Based on Gramian Angular Field and Coordinate Attention

Jialiang Cui, Qianwen Zhong *, Shubin Zheng *, Lele Peng and Jing Wen

School of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai 201620, China; sycuijialiang@icloud.com (J.C.); lele.peng@sues.edu.cn (L.P.); wenjing.jlu@126.com (J.W.)
* Correspondence: qianwen.zhong@sues.edu.cn (Q.Z.); shubin.zheng@sues.edu.cn (S.Z.)

Abstract: The key to ensuring rotating machinery’s safe and reliable operation is efficient and accurate faults diagnosis. Intelligent fault diagnosis technology based on deep learning (DL) has gained increasing attention. A critical challenge is how to embed the characteristics of time series into DL to obtain stable features that correlate with equipment conditions. This study proposes a lightweight rolling bearing fault diagnosis method based on Gramian angular field (GAF) and coordinated attention (CA) to improve rolling bearing recognition performance and diagnosis efficiency. Firstly, the time domain signal is encoded into GAF images after downsampling and segmentation. This method retains the temporal relation of the time series and provides valuable features for DL. Secondly, a lightweight convolution neural network (CNN) model is constructed through depthwise separable convolution, inverse residual block, and linear bottleneck layer to learn advanced features. After that, CA is employed to capture the long-range dependencies and identify the precise position information of the GAF images with nearly no additional computational overhead. The proposed method is tested and evaluated by CWRU bearing dataset and experimental dataset. The results demonstrate that the CNN based on GAF and CA (GAF-CA-CNN) model can effectively reduce the calculation overhead of the model and achieve high diagnostic accuracy.

Keywords: rolling bearing fault diagnosis; lightweight neural network; downsampling; gramian angular field; coordinate attention

1. Introduction

Rolling bearing is the critical component of rotating machinery and is widely used in rail transit, precision machine tools, aerospace, and other fields. Due to the constant impact of load, the rolling bearings are prone to cracks and pitting [1–3], which seriously affects equipment operation and even cause safety accidents and economic losses. Therefore, it is necessary to monitor the status of rolling bearings to ensure the regular operation of mechanical equipment.

The collision between the matching surface and the damage of the rolling bearing will produce a transient impact. If the rotation speed remains constant, it will produce periodic transient impact. Additionally, rolling bearings of different fault types have their specific vibration characteristics. Therefore, the fault diagnosis methods of rolling bearings are primarily based on the processing and analysis of vibration signals. The existing fault diagnosis methods for rolling bearings include signal processing-based and intelligent diagnosis methods. The former can effectively extract fault features from the original vibration signal, such as wavelet transform (WT) [4], empirical mode decomposition (EMD) [5], and variational mode decomposition (VMD) [6]. Li et al. [7] used an improved adaptive parameterless empirical wavelet transform (IAPEWT) for rolling bearing fault diagnosis. In 2019, Chen et al. [8] presented a rolling bearing fault diagnosis method based on EMD and sample quantile permutation entropy (SQPE). In 2020, Li et al. [9] designed a rolling bearing diagnosis model based on VMD and fractional Fourier transform (FRFT).
These methods rely on artificial feature extraction, which depends on excellent signal processing knowledge and engineering experience. With the development of computing technology, some studies combine signal processing and machine learning to diagnose failures. Qiao et al. [10] designed a rolling bearing fault detection model using the support vector machine (SVM) and improved EWT. Gunerker et al. [11] combined adaptive WT and K-nearest neighbor (KNN) algorithm to diagnose the bearing fault. These methods have been applied in practice, but the traditional machine learning methods are still complicated to extract the deep fault features from the original data.

In recent years, deep learning (DL) is applied in fault diagnosis and has excellent feature extraction ability which can automatically extract features from massive data. Additionally, due to the good performance in feature extraction, the intelligent diagnostic method for DL has been established. Wen et al. [12] proposed a bearing fault diagnosis model with 51 convolution layers using the method of transfer learning. Mao et al. [13] trained a bearing fault detection model based on VGG-16 and support vector data. Wang et al. [14] detected the bearing status of the inner wheel motor under different loads by an adaptive convolution neural network (CNN). In 2020, He et al. [15] designed an enhanced CNN structure to improve the performance of the diagnostic model for rotor bearings. Tian et al. [16] applied an improved deep CNN model framework to reduce the bearing fault detection error rate. Che et al. [17] designed a domain adaptive deep belief network (DBN) to realize fault diagnosis of rolling bearings under variable working conditions. Pang et al. [18] proposed an ensemble learning method to detect engine rolling bearing faults by denoising a multi-layer extreme learning machine.

CNN is widely used in computer vision because of its excellent feature extraction ability. There are some studies on converting one-dimensional vibration signals into two-dimensional images to make full use of the performance of CNN. Tao et al. [19] converted one-dimensional vibration signals of rolling bearings into two-dimensional images using the short-time Fourier transform (STFT). The experimental results show that this method has high diagnostic accuracy. Wang et al. [20] converts the bearing vibration signal into a 2D grayscale image and realize bearing fault diagnosis under different loads. These methods have improved bearing fault diagnosis performance to a certain extent. However, the challenge now is to embed the domain knowledge of rotating machinery into DL to obtain features related to the device’s health. Because of the periodicity of vibration signals, an image encoding method based on gramian angular field (GAF) [21] is presented in this study. GAF can preserve the temporal dependence of bearing vibration signals and change original data distribution, making it easier to distinguish from Gaussian noise.

However, traditional CNN structures, such as VGG-19 and Resnet-101, are usually bulky to obtain better diagnostic performance, resulting in efficiency problems. Hundreds of network layers mean extensive weights parameters, requiring high operating equipment and not meeting the actual application requirement. Therefore, this study proposed a method to process the downsampling of original vibration signal and construct a lightweight network structure via depthwise separable convolution [22] for reducing computational overhead. At the same time, the inverse residual structure and linear bottleneck layer [23] were introduced to improve the gradient propagation ability of the model and enhance the generalization performance of the model.

To further improve the performance of the model in practical applications, methods such as squeeze and excitation (SE) attention [24], bottleneck attention module (BAM) [25], and convolution block attention module (CBAM) [26] are proposed to introduce attention mechanism into the model. SE uses 2D global pooling to compute channel attention, significantly improving performance at a lower computational overhead. However, SE only considers channel information and ignores positional information, which is essential for capturing the structure of the image [27]. BAM and CBAM exploit positional information by reducing channel dimensions of input information and using convolution to calculate spatial attention. However, convolution can only capture local features, but it cannot learn the long-term dependencies of visual information [28]. Hou et al. proposed...
coordinate attention (CA) [29] in 2021. CA is a new attention mechanism that embeds location information into the channel attention, retaining the long-term dependencies and location information of visual information in different spatial directions while avoiding high computational cost. Therefore, CA is also used to enhance the feature extraction ability of the model in this study.

According to the requirements and characteristics of the rolling bearing fault diagnosis, a new fault diagnosis method for rolling bearings based on GAF-CA-CNN is proposed in this paper. First, the downsampling method is adopted to signal to implement data reduction, and the signal is divided into subsequences according to the rotation speed. Secondly, GAF is proposed to code the signal into a two-dimensional image. Finally, we train the lightweight CA-CNN model to identify rolling bearing failure types.

The main contributions of this study are as follows: (1) The proposed image encoding method can embed the temporal correlation of the vibration signal into the visual representation and change the distribution of the data so that it can be easily separated from Gaussian noise. (2) Construct a lightweight CNN model to reduce the operating costs so that the model can meet the practical diagnostic needs. (3) Using CA enables the model to focus on important information, which improves the learning ability of the model.

This paper is organized into five parts. After the introduction, Section 2 investigates the principles of the algorithm. Section 3 describes the fault diagnosis process of the proposed method. The analysis and discussion of experimental validation is arranged in Section 4. Section 5 is the conclusion.

2. Theoretical Background

2.1. Gramian Angular Field (GAF)

In rotating machinery, the bearing vibration signal is periodic. The random noise tends to impact the periodic vibration signal, so it is difficult to extract the bearing fault features directly from the time-domain signal. GAF provides a method to encode time-domain signals into images, separating characteristic signals from interference signals while preserving the temporal relationship of signals. At present, GAF has achieved some results in human behavior recognition (HAR) [30], solar radiation prediction [31], and ECG signal monitoring [32].

For a given time series \( X = x_1, x_2, \ldots, x_n \), normalize and scale \( X \) to \([-1,1]\) by the Equation (1).

\[
\tilde{x}_i = \frac{x_i - \max(X) + (x_i - \min(X))}{\max(X) - \min(X)} \quad (1)
\]

The angle \( \varphi \) is the inverse cosine of \( x_i \), the radius \( r \) is the timestamp, and convert the time series \( X \) into polar coordinates [20].

\[
\begin{cases}
\varphi = \arccos(\tilde{x}_i), & -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \mathbb{X} \\
\quad r = \frac{t_i}{N}, t_i \in \mathbb{N}
\end{cases} \quad (2)
\]

From Equation (2), the \( t_i \) is the timestamp. \( N \) is used as a constant to adjust the span of image torsion on polar coordinates. The time series will show a “diffusion” in polar coordinates as time increases. Generally, a mapping that is both injective and surjective is called bijection. From Equation (2), when \( \varphi \in [0, \pi] \), the \( \cos(\varphi) \) is monotonic and bijective. That means, given any time series, the proposed method produces only one result in the polar coordinate system with a unique inverse function. Moreover, unlike Cartesian coordinates, the polar coordinates preserve absolute temporal relation.

Figure 1a–d shows the process of converting time-domain signals into GAF images.
The Gram matrix is composed of the inner product of any \( k(k \leq n) \) vectors \( \alpha_1, \alpha_2 \ldots \alpha_k \) in \( n \)-dimensional Euclidean space.

\[
\Delta(\alpha_1, \alpha_2, \ldots, \alpha_k) = \begin{bmatrix}
(\alpha_1, \alpha_1) & (\alpha_1, \alpha_2) & \cdots & (\alpha_1, \alpha_k) \\
(\alpha_2, \alpha_1) & (\alpha_2, \alpha_2) & \cdots & (\alpha_2, \alpha_k) \\
\vdots & \vdots & \ddots & \vdots \\
(\alpha_k, \alpha_1) & (\alpha_k, \alpha_2) & \cdots & (\alpha_k, \alpha_k)
\end{bmatrix}
\]  

(3)

The Gram matrix is used to measure the characteristics and the relationship between each dimension. In the multi-scale matrix obtained after the inner product, the main diagonal element provides information about the feature itself. In contrast, the other elements reflect the relevant information between different features. After transforming the time series into the polar coordinate system, the Gram matrix calculates the temporal relation in different time intervals. Define Gramian angular field \( G \) as follows [21]:

\[
G = \begin{bmatrix}
\cos(\phi_1 + \phi_1) & \cdots & \cos(\phi_1 + \phi_n) \\
\cos(\phi_2 + \phi_1) & \cdots & \cos(\phi_2 + \phi_n) \\
\vdots & \ddots & \vdots \\
\cos(\phi_n + \phi_1) & \cdots & \cos(\phi_n + \phi_n)
\end{bmatrix}
\]  

(4)

From Equation (4): \( \cos(\phi_1 + \phi_2) = \cos(\arccos(\phi_1) + \arccos(\phi_2)) \), according to Equation (5)

\[
\cos(x + y) = \cos(x) \cdot \cos(y) - \sin(x) \cdot \sin(y)
\]  

(5)

\[
\cos(\phi_1 + \phi_2) = \cos(\arccos(\phi_1) + \arccos(\phi_2)) = \cos(\arccos(\phi_1)) \cdot \cos(\arccos(\phi_2)) - \sin(\arccos(\phi_1)) \cdot \sin(\arccos(\phi_2))
\]  

(6)
Define \( u = \cos(\arccos(x_1)) \), \( v = \cos(\arccos(x_2)) \), we can define a new inner product:

\[
\langle u, v \rangle = \sum_{i=1}^{n} u_i \cdot v_i - \sqrt{1 - u_i^2} \cdot \sqrt{1 - v_i^2}
\]  

(7)

The new inner product has a penalty item \( \sqrt{1 - u_i^2} \cdot \sqrt{1 - v_i^2} \) compared with \( \langle u, v \rangle = \sum_{i=1}^{n} u_i \cdot v_i \). Figure 2a–c shows the effect of the penalty item.

![Figure 2](image)

**Figure 2.** The process of constructing cosine inner product, (a) Standard inner product density distribution and 3d image, (b) Penalty density distribution and 3d image, (c) Cosine inner product density distribution and 3d image.

Figure 2a shows that the original inner product Gram matrix follows a Gaussian distribution centered on 0. The more Gaussian the distribution of data, the more difficult it is to distinguish it from Gaussian noise. Figure 2b shows that the penalty will be more significant when the point is closer to 0. These points are closer to Gaussian noise. Figure 2c shows that the density distribution of the output becomes non-sparse and easy to separate with Gaussian noise.

GAF provides a method to maintain time correlation. Time increases along the main diagonal of the Gram matrix. \( G_{i(j)l(j)=k} \) represents the correlation of time series at the time interval \( k \). The main diagonal element \( G_{i(j)} \) is composed of the original values of the scaled time series. However, the length of the main diagonal of GAF is \( n^2 \) and the original time series is \( n \). The Piecewise Aggregation Approximation (PAA) [33] is proposed to smooth the time series.

2.2. Coordinate Attention (CA)

In this sub-section, CA [29] is introduced to improve the convergence speed and test the model’s accuracy. The CA has the following two advantages:

1. It can capture the characteristics of different channels. The model can identify targets more accurately by capturing direction-aware and position-aware information.
2. The CA module is light and flexible enough to be inserted into various models easily.

Figure 3 shows the structure of the CA. Each channel encodes vertical and horizontal coordinates for the characteristic graph \( X \in \mathbb{R}^{H \times W \times C} \) using pooling kernels \((1, W)\) and \((H, 1)\), respectively. The unidirectional pooling result of the \( c-th \) channel at height \( h \) can be formulated as [29]:

\[
z_{c}^{h}(h) = \frac{1}{W} \sum_{0 \leq i < W} x_{c}(h, i)
\]

(8)
Similarly, the output of the $c$–th channel at weight $w$ can be formulated as [29]:

$$ z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \quad (9) $$

The above two transformations aggregate features along with the two spatial directions, respectively, yielding a pair of the direction-aware feature map. These two transformations also capture the long-range dependencies alone in one spatial direction and preserve precise positional information along the other spatial direction, which helps the network locate the object of interest more accurately.

The two generated feature maps are combined in the same direction and $1 \times 1$ convolution to extract their features, and the generated feature map will contain spatial information in the horizontal and vertical directions, which is expressed as follows [29]:

$$ f = \text{ffi}\left(F_{1}\left(\left[z_h^w, z_w^w\right]\right)\right) \quad (10) $$

The feature map $f \in \mathbb{R}^{C/h \times (H+W)}$ is cut into two different tensors $f^h$ and $f^w$ along the spatial dimension, raising the dimensions of tensors by $1 \times 1$ convolution $F_{1}$ transformation to obtain the same dimension as the input $X \in \mathbb{R}^{H \times W \times C}$. $\text{ffi}$ is the Non-linear activation function. The formula is as follows [29] to calculate the attention weight:

$$ g_h^h = \sigma(F_h(f^h)) \quad (11) $$

$$ g_w^w = \sigma(F_w(f^w)) \quad (12) $$

$\sigma$ is Sigmoid function and $F_{h,w}$ is $1 \times 1$ convolution. Finally, the output of the CA $V_c$ is expressed as [29]:

$$ V_c(i, j) = X_c(i, j) \times g_h^h(i, j) \times g_w^w(i, j) \quad (13) $$

3. The Proposed Method

The occurrence of bearing fault will produce a series of periodic shocks. In industrial machinery, the noise signal may mask the impact signal, resulting in the inability to obtain practical features directly from the time-domain signal. After downsampling, the vibration signals are divided into subsequences to reduce the amount of data under high frequency. GAF is introduced to process the subsequence samples to obtain a superior feature representation. Deep separable convolution, inverse residual module, and linear bottleneck layer build the CNN framework. The CA is inserted into the CNN framework to augment the representations of the GAF of interest.
3.1. Data Downsampling

To overcome the shortcoming of extensive computation, the downsampling method is introduced to the original vibration data. However, the model must consider the Nyquist sampling theorem before downsampling, which stipulates that the sampling rate shall not be less than twice the maximum frequency of the signal. For example, when the 48 kHz signal is downsampled to 16 kHz, the signal will have a Nyquist frequency of 8 kHz, which means that the spectral components at 9 kHz and 7 kHz will become a challenge to the downsampled signal. The reason is that the 8 kHz is mixed into the 9 kHz signal. The research in [34] indicates that the spectrum folding is consistent with no low-pass filter in the downsampling process. Hence, in order to suppress signal aliasing, a digital low-pass filter should be used to attenuate the frequency component above 8 kHz for the 48 kHz signal before downsampling.

In this paper, the resample, fil1, and Kaiser commands are used in MATLAB® (MathWorks, Inc., Natick, MA, USA) to perform downsampling. The resample instruction uses an anti-aliasing filter and adjusts the downsampling process by the ratio of the original signal to the target signal frequency. However, artifacts are brought into the downsampled signal in this process, so it is necessary to set a roll-off anti-aliasing filter to generate a spectrum gap to replace the aliasing artifacts. Algorithm 1 shows the process.

Algorithm 1: Design Kaiser window to approximate the anti-aliasing filter

| Input: original data frequency q, downsample data frequency p, original data |
| Output: downsampling data |
| 1 Cutoff frequency $f_c = 1/\max(p,q)$ |
| 2 Define the parameter $n$ to control the roll-off band |
| 3 Define shape parameter beta to control the tradeoff between transition width and stopband attenuation |
| 4 If $p/q \neq$ integer then |
| 5 Insert zeros to upsampling the signal by $q$ |
| 6 filter order = $2 \times n \times \max(p,q)$ |
| 7 filter = fil1(filter order, $f_c$, Kaiser(filter order + 1, beta)) |
| 8 Apply an anti-aliasing filter to the upsampling data |
| 9 Discard samples to downsampling the filtered signal by $p$ |
| 10 Final |
| 11 Return |

3.2. The Methods for Lightweight Network

3.2.1. Depthwise Separable Convolution

The lightweight model reduces the number of model parameters by decomposing convolution kernel and singular value decomposition to speed up the network calculation [22]. Common lightweight models include Mobilenet [35], SqueezeNet [36], Xception [37], and Sufflenet [38]. The four models compress parameters differently to realize a lightweight network, effectively reducing model parameters and retaining good accuracy.

As an alternative to traditional convolution, Deepwise separation convolution [22] is widely used in lightweight models, as shown in Figure 4. An $N \times N$ standard convolution of depthwise separable convolution is decomposed into an $N \times N$ Depthwise convolution (DW) and a $1 \times 1$ Pointwise Convolution (PW). The former function uses one convolution kernel for each input channel, while the latter combines the results to ensure that the input and output have the same size.
6 \text{ filter order} = 2 \times n \times \max (p, q) 

7 \text{ filter} = \text{fir1} (\text{filter order}, ... 

Figure 4. Standard convolution and depthwise separable convolution.

It is assumed that the standard convolution has \( k \) sizes of \( N \times N \) channel, and the input image size is \( H \times W \times C \). Then the computational overhead of the standard convolution layer is:

\[
N \times N \times K \times C \times H \times W
\]  

(14)

The computational overhead of the depthwise separable convolution is:

\[
\frac{N \times N \times C \times H \times W + C \times K \times H \times W}{N \times N \times K \times C \times H \times W} = \frac{1}{K} + \frac{1}{N^2}
\]  

(15)

As depthwise separable convolution uses more than ten filters of size \( 3 \times 3 \) typically, the computational overhead by Equation (14) is more than that calculated by Equation (15).

3.2.2. Inverted Residual Block with Linear Bottleneck

Each convolution layer uses ReLU as the activation function in the general DL model. However, the output of neurons in the deep convolution layer is easy to approach 0 due to the reduction of the number of extracted features when the convolution kernel with a lower dimension is input. The ReLU will probably lead to zero tensors in a particular dimension under low dimensional tensors, resulting in irreversible information loss. Therefore, the reverse residual block with a linear bottleneck layer [23] is introduced to directly take the output of the convolution layer as the input of the next layer.

The inverse residual block is composed of three-layer convolution. Firstly, the bottleneck layer enhances the channel dimension, and then the deep convolution layer is used to extract the features. Then, the linear bottleneck layer is applied again to map the extracted features in the low dimensional space for reducing the loss of information. Figure 5 shows the structure of the standard residual block and the reverse residual block. The output of the ReLU activation function in the reverse residual block is limited to 6.
The computational overhead of GAP is much less than that of the FC layer.\[ H \cdot W \cdot C \cdot n \] 

Figure 5. Residual block and inverted residual block, (a) Residual block, (b) Inverted residual block with linear Bottleneck.

3.2.3. Global Average Pooling (GAP)

The traditional convolutional neural network uses the full connection layer (FC) and softmax classifier to output the prediction results of the model. However, the FC layer requires many training and tuning parameters, which reduces the training speed and is easy to produce overfitting. Therefore, using global average pooling (GAP) [39] to pool the feature map of each layer not only reduces the number of parameters but also better corresponds to the channel to the feature map. If the category required by the classifier is \( n \) and the feature size is \( H \times W \times C \) then the computational overhead of the parameters of the FC layer is \( H \cdot W \cdot C \cdot n \) and the computational overhead of GAP is \( 1 \cdot C \cdot n \). The computational overhead of GAP is much less than that of the FC layer.

Figure 6 shows the structure of GAP. GAP performed an average of each feature map, and the resulting vector is fed directly into the softmax layer. This strategy is more native to the convolution structure by enforcing correspondences between feature maps and categories. Moreover, there is no parameter to optimize in GAP and avoid overfitting.

Figure 6. Global average pooling.

3.3. General Procedures of the Proposed Method

This section proposes combining CNN and GAF for rolling bearing fault diagnosis. Figure 7 shows the flow chart of the proposed method. The proposed method includes three main steps.
Figure 7. The framework of the proposed method.

Step 1: the vibration signals, operating under different rolling bearings, are collected and downsampled to a lower frequency. Then, the signal is divided into segments to encode GAF images, which are resized to 64 × 64 size as the model’s input.

Step 2: Multiple reverse residual blocks stacked the CNN framework with a linear bottleneck layer. The deep separable convolution replaces the regular convolution to reduce the computational overhead. GAP is proposed to replace the traditional FC layer in CNN, which provides a robust spatial transition of the input.

Step 3: the image is randomly divided into training and test samples. The fault diagnosis experiment is carried out in the GAF-CA-CNN model and optimized by the Adam algorithm [40], which adapts the learning rate and improves the training speed. Finally, classification and visualization results are given to provide a comprehensive diagnostic analysis.
4. Validation and Discussion

4.1. Case 1: CWRU Bearing Dataset

This section applies the GAF-CA-CNN model to the bearing data set of Case Western Reserve University Laboratory for verification [41]. To verify the performance of the proposed method in dealing with the diversity of bearing faults, we took ten bearing states under 0 load and tested them, including nine bearing faults and one normal state, as shown in Table 1. The dataset includes 10,000 image samples, which are randomly divided into the training set, verification set, and test set according to 7:2:1.

| Label | Fault Size (mm) | States                  | Motor Speed (r/min) | Sample Size            |
|-------|-----------------|-------------------------|---------------------|------------------------|
| 1     | —               | Normal (N)              | 1797                | 1000 × 642 × 3         |
| 2     | 0.1778          | Inner race (IF1)        | 1797                | 1000 × 642 × 3         |
| 3     | 0.1778          | Roll boll (RF1)         | 1797                | 1000 × 642 × 3         |
| 4     | 0.1778          | Outer race (OF1)        | 1797                | 1000 × 642 × 3         |
| 5     | 0.3556          | Inner race (IF2)        | 1797                | 1000 × 642 × 3         |
| 6     | 0.3556          | Roll boll (RF2)         | 1797                | 1000 × 642 × 3         |
| 7     | 0.3556          | Outer race (OF2)        | 1797                | 1000 × 642 × 3         |
| 8     | 0.5334          | Inner race (IF3)        | 1797                | 1000 × 642 × 3         |
| 9     | 0.5334          | Roll boll (RF3)         | 1797                | 1000 × 642 × 3         |
| 10    | 0.5334          | Outer race (OF3)        | 1797                | 1000 × 642 × 3         |

4.1.1. Environment Setup

The parameter settings of GAF-CA-CNN are shown in Table 2 where the Up represents the dimension raised in the Bottleneck. HS indicates the Hard-swish activation function as shown in Equation (16). RE indicates ReLU6. All tests were carried out on a computer with AMD R7000 2.9 GHz CPU and 16 g RAM. GAF-CA-CNN is based on the Tensorflow2-GPU framework of Python 3.7. Figure 8 shows the training process of the model under different attention mechanisms. The model with the CA module has better training efficiency and recognition accuracy.

\[
h - \text{swish}(x) = x \frac{\text{ReLU6}(x + 3)}{6}
\] (16)

In this manuscript, the GAP method is applied to replace the full connected (FC) layer, thus reducing the model training parameters, and GAP corresponds to GMP (Global Maximum Pooling). Table 3 shows the comparison results.

| Input         | Module        | Up    | Output | Attention | Activation | Sample Size |
|---------------|---------------|-------|--------|-----------|------------|-------------|
| 64\(^2\) × 3  | Conv_block    | —     | 16     | —         | HS         | 2           |
| 32\(^2\) × 16 | Bottleneck    | 16    | 16     | True      | RE         | 2           |
| 16\(^2\) × 16 | Bottleneck    | 36    | 24     | False     | RE         | 2           |
| 8\(^2\) × 24  | Bottleneck    | 44    | 24     | False     | RE         | 1           |
| 8\(^2\) × 24  | Bottleneck    | 48    | 40     | True      | HS         | 2           |
| 4\(^2\) × 40  | Bottleneck    | 120   | 40     | True      | HS         | 1           |
| 4\(^2\) × 40  | Bottleneck    | 120   | 40     | True      | HS         | 1           |
| 4\(^2\) × 40  | Bottleneck    | 60    | 48     | True      | HS         | 1           |
| 4\(^2\) × 48  | Bottleneck    | 72    | 48     | True      | HS         | 1           |
| 4\(^2\) × 48  | Bottleneck    | 144   | 96     | True      | HS         | 2           |
| 2\(^2\) × 96  | Bottleneck    | 288   | 96     | True      | HS         | 1           |
| 2\(^2\) × 96  | Bottleneck    | 288   | 96     | True      | HS         | 1           |
| 2\(^2\) × 96  | Conv_block    | —     | 288    | —         | HS         | 1           |
| 1\(^2\) × 288 | Glob_avg_pool | —     | —      | —         | —          | —           |
| 1\(^2\) × 512 | Conv2d        | —     | 512    | —         | HS         | 1           |
| 1\(^2\) × 10  | Conv2d        | —     | 10     | —         | Softmax    | 1           |
6.0.3556 Roll boll (RF2) 1797 1000 × 642 × 3
7.0.3556 Outer race (OF2) 1797 1000 × 642 × 3
8.0.5334 Inner race (IF3) 1797 1000 × 642 × 3
9.0.5334 Roll boll (RF3) 1797 1000 × 642 × 3

4.1.1. Environment Setup
The parameter settings of GAF-CA-CNN are shown in Table 2 where the Up represents the dimension raised in the Bottleneck. HS indicates the Hard-swish activation function as shown in Equation (16). RE indicates ReLU6. All tests were carried out on a computer with AMD R7000 2.9 GHz CPU and 16 g RAM. GAF-CA-CNN is based on the Tensorflow2-GPU framework of Python 3.7. Figure 8 shows the training process of the model under different attention mechanisms. The model with the CA module has better training efficiency and recognition accuracy.

Table 2. Details of GAF-CA-CNN.

| Input Module                  | Up      | Output | Attention | Activation | Sample Size |
|-------------------------------|---------|--------|-----------|------------|-------------|
| 64 × 2 × 3 Conv_block         | —       | 16     | —         | HS         | 2           |
| 32 × 2 × 16 Bottleneck        | 16      | 16     | True      | RE         | 2           |
| 16 × 2 × 16 Bottleneck        | 36      | 24     | False     | RE         | 2           |
| 8 × 2 × 24 Bottleneck         | 44      | 24     | False     | RE         | 1           |
| 8 × 2 × 24 Bottleneck         | 48      | 40     | True      | HS         | 2           |
| 4 × 2 × 40 Bottleneck         | 120     | 40     | True      | HS         | 1           |
| 4 × 2 × 40 Bottleneck         | 120     | 40     | True      | HS         | 1           |
| 4 × 2 × 40 Bottleneck         | 60      | 48     | True      | HS         | 1           |
| 4 × 2 × 48 Bottleneck         | 72      | 48     | True      | HS         | 1           |
| 2 × 2 × 96 Bottleneck         | 120     | 96     | True      | HS         | 2           |
| 2 × 2 × 96 Bottleneck         | 288     | 96     | True      | HS         | 1           |
| 2 × 2 × 96 Conv_block         | —       | 288    | —         | HS         | 1           |
| 1 × 2 × 512 Conv2d            | —       | 512    | —         | HS         | 1           |
| 1 × 2 × 10 Conv2d             | —       | 10     | —         | Softmax    | 1           |

(a) (b)

Figure 8. Present the training process of the model with different attention modules. (a) Loss, (b) Accuracy.

Due to the global average of GAP, GAP loss drives the network to distinguish each category, which can be the prediction for finding all target distinguishable areas. Additionally, GMP is maximized globally—only the region with the highest score can be found and other regions with low scores would be ignored.

Figure 9 shows the comparison between the original signal and the downsampled signal. Downsampling from 12 kHz to 10 kHz reduces data by 16.7%, while that from 12 kHz to 8 kHz reduces data by 33.3%. Moreover, downsampling from 12 kHz to 6 kHz reduces data by 50%, and that from 12 kHz to 4 kHz is followed by 66.7%. Compared with those results, downsampling from 12 kHz to 2 kHz reduces data by 83.3% and 12 kHz~1 kHz reduces data by 91.6%, respectively.

As the amount of data is reduced, less information is included in each GAF image. Thus, controlling the number of images is necessary to compensate for the lack of information. We set 1000 images per rolling bearing conditions in different frequencies. Figures 10 and 11 show the accuracy and loss in different sampling frequencies. The convergence of the model is best when the sampling rate is 2 kHz. At the same time, with the increase in sampling rate, fluctuation of the model loss greatly increases. There are many vibration data points at a high sampling rate, and more images are needed to represent effectively represent the signal features. One thousand images cannot meet the number of features required by the model at a high sampling rate. Figure 12 shows the average accuracy and loss at each sampling frequency, and 2 kHz has the best performance.
The GAP method is applied to replace the full connected (FC) layer, thus reducing the model training parameters, and GAP corresponds to GMP (Global Maximum Pooling). Table 3 shows the comparison results.

Table 3. The average loss of Global Average Pool and Global Maximum Pool.

| Methods        | Average Loss 1 | Average Loss 2 | Average Loss 3 |
|----------------|----------------|----------------|----------------|
| GAF-CA-CNN-GAP | 0.140          | 0.194          | 0.158          |
| GAF-CA-CNN-GMP | 0.169          | 0.237          | 0.182          |

Due to the global average of GAP, GAP loss drives the network to distinguish each category, which can be the prediction for finding all target distinguishable areas. Additionally, GMP is maximized globally—only the region with the highest score can be found and other regions with low scores would be ignored.

Figure 9 shows the comparison between the original signal and the downsampled signal. Downsampling from 12 kHz to 10 kHz reduces data by 16.7%, while that from 12 kHz to 8 kHz reduces data by 33.3%. Moreover, downsampling from 12 kHz to 6 kHz reduces data by 50%, and that from 12 kHz to 4 kHz is followed by 66.7%. Compared with those results, downsampling from 12 kHz to 2 kHz reduces data by 83.3% and 12 kHz~1 kHz reduces data by 91.6%, respectively.

As the amount of data is reduced, less information is included in each GAF image. Thus, controlling the number of images is necessary to compensate for the lack of information. We set 1000 images per rolling bearing conditions in different frequencies. Figures 10 and 11 show the accuracy and loss in different sampling frequencies. The convergence of the model is best when the sampling rate is 2 kHz. At the same time, with the increase in sampling rate, fluctuation of the model loss greatly increases. There are many vibration data points at a high sampling rate, and more images are needed to represent effectively represent the signal features. One thousand images cannot meet the number of features required by the model at a high sampling rate. Figure 12 shows the average accuracy and loss at each sampling frequency, and 2 kHz has the best performance.
4.1.2. Experimental Result and Analysis

This section verifies the advantages of the proposed method in model size and prediction accuracy. Figure 13 shows that after the timing signal is encoded into a GAF image, the GAF images of each state of the bearing are different. The main diagonal of the GAF image represents the time change. The highlighted lines in the horizontal and vertical directions show that the amplitude changes sharply and quickly.

Table 4 shows the size and prediction speed of different models. The number of GAF-CA-CNN parameters is lower than the traditional DL and machine learning models and has higher diagnostic efficiency. Even if the results are similar, the size of the proposed method is 1/40 of VMD-Gray image-Resnet50 and EMD-Gray image-Resnet50.

Table 4. Parameter size and prediction speed of different models.

| Models           | Model Size (MB) | Prediction Speed (ms) |
|------------------|-----------------|-----------------------|
| GAF-CBAM-CNN     | 2.26 x 10^4     |                       |
| GAF-BAM-CNN      | 18.7 x 113      |                       |
| VMD-Gray image-Resnet50 | 90.5 x 71    |                       |
| EMD-Gray image-Resnet50 | 90.5 x 78    |                       |

Figure 11. The validation loss of Dataset A (a) 1 kHz, 2 kHz, 4 kHz, and 6 kHz, (b) 8 kHz, 10 kHz, and 12 kHz.

Figure 12. The average accuracy and loss of different frequencies on dataset A.

Figure 13. The GAF images of Dataset A.
and has higher diagnostic efficiency. Even if the results are similar, the size of the proposed method is 1/40 of VMD-Gray image-Resnet50 and EMD-Gray image-Resnet50.

**Table 4.** Parameter size and prediction speed of different models.

| Models                        | Model Size (MB) | Prediction Speed (ms) |
|-------------------------------|-----------------|-----------------------|
| GAF-CA-CNN                    | 2.20            | 60                    |
| GAF-SE-CNN                    | 2.37            | 55                    |
| GAF-CBAM-CNN                  | 2.26            | 104                   |
| GAF-BAM-CNN                   | 18.7            | 113                   |
| VMD-Gray image-Resnet50       | 90.5            | 71                    |
| EMD-Gray image-Resnet50       | 90.5            | 78                    |

The dataset at 2 kHz in dataset A was tested five times. Table 5 and Figure 14 describe the average diagnostic rates of different models. It can be seen from the figure that the average diagnostic accuracy of the proposed method is 99.62%, which is 5.89% higher than that of the SE attention model, 0.77% higher than that of the CBAM (Convolutional Block Attention Module) model, 1.39% higher than the BAM (Bottleneck Attention Module) model, 9.27% higher than that of the VMD Gray image-Resnet50 model, and 3.29% higher than that of EMD-Gray image-Resnet50. The last two methods apply EMD (Empirical Modal Decomposition) and VMD (Variational Modal Decomposition) to extract IMF components of signals and recode them into gray images.

**Table 5.** Diagnostics result of different methods on Dataset A.

| Models                        | 1       | 2       | 3       | 4       | 5       | Average | Standard Deviation |
|-------------------------------|---------|---------|---------|---------|---------|---------|-------------------|
| GAF-CA-CNN                    | 99.70%  | 99.55%  | 99.75%  | 99.35%  | 99.75%  | 99.62%  | 0.154             |
| GAF-SE-CNN                    | 98.90%  | 89.70%  | 96.05%  | 84.50%  | 99.50%  | 93.73%  | 5.776             |
| GAF-CBAM-CNN                  | 98.85%  | 99.25%  | 97.70%  | 98.70%  | 99.75%  | 98.85%  | 0.680             |
| GAF-BAM-CNN                   | 99.05%  | 96.05%  | 98.45%  | 99.65%  | 97.95%  | 98.23%  | 1.230             |
| VMD-Gray image-Resnet50       | 91.25%  | 91.75%  | 94.75%  | 87.50%  | 86.50%  | 90.35%  | 3.002             |
| EMD-Gray image-Resnet50       | 95.33%  | 94.67%  | 95.00%  | 98.33%  | 98.33%  | 96.33%  | 1.637             |

The results show that compared with other methods, GAF-CA-CNN can better learn the characteristics of the original signal, and CA can enhance feature learning and stabilize the training process.

Figure 15 shows the t-SNE visualization of diagnostic results of different methods. The characteristics of multiple states of VMD and EMD methods are overlapped. The characteristics of RF1 and RF3 are overlapped in the CBAM method. SE method and BAM
method can clearly distinguish a variety of bearing states. The results show that GAF-CA-CNN can automatically extract practical features and realize fast, high-precision fault diagnoses. Figure 16 shows the confusion matrix of these methods and summarizes the diagnostic results of different methods. The horizontal axis is the predicted bearing state, and the vertical axis is the actual bearing state. The data on the main diagonal represents the accuracy of the corresponding bearing state prediction. The VMD method and EMD method have prediction errors in multiple bearing states. The other method can maintain high diagnostic accuracy.

![Confusion Matrix](image)

**Figure 16.** The confusion matrix of different diagnostic results on dataset A, (a) The proposed method, (b) GAF-SE-CNN, (c) GAF-CBAM-CNN, (d) GAF-BAM-CNN, (e) VMD-Gray image-Resnet50, (f) EMD-Gray image-Resnet50.

### 4.2. Case 2: Laboratory Dataset

#### 4.2.1. Experiment Preparation

To further verify the effectiveness of GAF-CA-CNN in vibration signal feature learning, bearing fault diagnosis is carried out on the bearing test-bed. The test bench configuration and bearing structure are shown in Figure 17, and are composed of a motor, transmission shaft, bearing seat, load application device, and pressure sensor. Table 6 shows the parameters of SKF 6016-2RS1. The Wavebook516E wired acquisition instrument of IOTECH company was used to record the vibration signal. The piezoelectric accelerometer is 1A110E type of Donghua Test and the motor is 1LE0001-1AA4 with a maximum speed is 2885 r/min. The pressure sensor is BSCC-ZN4S type and it shows the force applied to the rolling bearing. The sampling frequency was 10 kHz, and the rotation rate was 540 r/min. The dataset includes one normal state, four kinds of the inner ring, four kinds of the outer ring, and one fault state of the roller in Table 7. Dataset B is divided into the training set, verification set, and test set according to 7:2:1.
4.2. Case 2: Laboratory Dataset

To further verify the effectiveness of GAF-CA-CNN in vibration signal feature learning, bearing fault diagnosis is carried out on the bearing test-bed. The test bench configuration and bearing structure are shown in Figure 17, and are composed of a motor, transmission shaft, bearing seat, load application device, and pressure sensor. Table 6 shows the parameters of SKF 6016-2RS1. The Wavebook516E wired acquisition instrument of IOTECH company was used to record the vibration signal. The piezoelectric accelerometer is 1A110E type of Donghua Test and the motor is 1LE0001-1AA4 with a maximum speed of 2885 r/min. The pressure sensor is BSCC-ZN4S type and it shows the force applied to the rolling bearing. The sampling frequency was 10 kHz, and the rotation rate was 540 r/min. The dataset includes one normal state, four kinds of the inner ring, four kinds of the outer ring, and one fault state of the roller in Table 7. Dataset B is divided into the training set, verification set, and test set according to 7:2:1.

Table 6. The parameters of SKF 6016-2RS1.

| Parameter | Size (mm) |
|-----------|-----------|
| d         | 80        |
| D         | 125       |
| B         | 22        |
| d1        | ≈94.4     |
| D2        | ≈114.1    |
| r1,2      | Min. 1.1  |
Table 7. Dataset B: Experiment bearing operation states.

| Label | Fault Size (mm) | States       | Motor Speed (r/min) | Sample Size       |
|-------|-----------------|--------------|---------------------|-------------------|
| 1     | —               | Normal (N)   | 540                 | 1000 × 642 × 3    |
| 2     | Abrasion        | Roll boll (RF1) | 540 | 1000 × 642 × 3 |
| 3     | Single column pitting | Inner race (IF1) | 540 | 1000 × 642 × 3 |
| 4     | Single column pitting | Outer race (OF1) | 540 | 1000 × 642 × 3 |
| 5     | Double column pitting | Inner race (IF2) | 540 | 1000 × 642 × 3 |
| 6     | Double column pitting | Outer race (OF2) | 540 | 1000 × 642 × 3 |
| 7     | 3               | Inner race (IF3) | 540 | 1000 × 642 × 3 |
| 8     | 3               | Outer race (OF3) | 540 | 1000 × 642 × 3 |
| 9     | 6               | Inner race (IF4) | 540 | 1000 × 642 × 3 |
| 10    | 6               | Outer race (OF4) | 540 | 1000 × 642 × 3 |

Figures 18 and 19 show the downsampling result of dataset B. The 1 kHz has the best convergence speed and effect. However, the number of images cannot fully represent the characteristics of the original signal under high sampling rate, the diagnosis curve fluctuates sharply, and the model cannot complete convergence in 200 generations. Figure 20 shows the average accuracy and loss at each sampling frequency. The model has the best performance at 1 kHz.

Figure 18. The validation accuracy of dataset B, (a) 1 kHz, 2 kHz, 4 kHz, (b) 5 kHz, 6 kHz, 8 kHz, and 10 kHz.

Figure 19. The validation loss of dataset B, (a) 1 kHz, 2 kHz, 4 kHz, (b) 5 kHz, 6 kHz, 8 kHz, and 10 kHz.
Figures 18 and 19 show the downsampling result of dataset B. The 1 kHz has the best convergence speed and effect. However, the number of images cannot fully represent the characteristics of the original signal under high sampling rate, the diagnosis curve fluctuates sharply, and the model cannot complete convergence in 200 generations. Figure 20 shows the average accuracy and loss at each sampling frequency. The model has the best performance at 1 kHz.

Figure 18. The validation accuracy of dataset B, (a) 1 kHz, 2 kHz, 4 kHz, (b) 5 kHz, 6 kHz, 8 kHz, and 10 kHz.

Figure 19. The validation loss of dataset B, (a) 1 kHz, 2 kHz, 4 kHz, (b) 5 kHz, 6 kHz, 8 kHz, and 10 kHz.

Figure 20. The accuracy and loss of dataset B.

4.2.2. Experimental Result and Analysis

This section is to verify the model’s performance on dataset B. Figure 21 shows the GAF image of each state of the bearing at 1 kHz. Compared with dataset A, the highlighted lines in the GAF image of dataset B are thinner due to the different rotational speeds of the two datasets, resulting in different sub-sample lengths of the generated GAF image. According to Equation (17):

\[ n = \frac{f}{\omega/60} \]  

(17)

The calculation shows that \( n_A = 222 \) at the 2 kHz sampling rate in dataset A and \( n_B = 111 \) at the 1 kHz sampling rate in dataset B. The GAF image generated by dataset B contains an enormous amount of data. Therefore, the corresponding highlighted lines are thinner in dataset B.

Similarly to dataset A, the data at 1 kHz in dataset B is also tested five times. Table 8 and Figure 22 show that the diagnostic accuracy of the proposed method is 99.91%. The accuracy is 1.5% higher than the SE model, 2.93% higher than the CBAM, and 2.92% higher than the BAM model. The accuracy of the VMD model and EMD model is 71.20% and 93.73%.

Table 8. Diagnostics result of different methods on Dataset B.

| Models                       | 1    | 2    | 3    | 4    | 5    | Average | Standard Deviation |
|------------------------------|------|------|------|------|------|---------|-------------------|
| GAF-CA-CNN                   | 99.9 | 98.75| 98.5 | 98.1 | 94.75| 94      | 95.6              |
| GAF-SE-CNN                   | 99.9 | 98.75| 98.5 | 98.1 | 94.75| 94      | 95.6              |
| GAF-CBAM-CNN                 | 94.1 | 98.4 | 96.2 | 99.3 | 96.9 | 96.98   | 1.8059            |
| GAF-BAM-CNN                  | 99.8 | 99.3 | 93.4 | 98.6 | 93.7 | 96.99   | 2.8176            |
| VMD-Gray image-Resnet50      | 67.75| 75.25| 68.5 | 70.0 | 74.5 | 71.20   | 3.0959            |
| EMD-Gray image-Resnet50      | 96.7 | 95.3 | 93.0 | 87.3 | 96.3 | 93.73   | 3.4484            |

Figure 21. The GAF images of dataset B.

The calculation shows that \( n_A = 222 \) at the 2 kHz sampling rate in dataset A and \( n_B = 111 \) at the 1 kHz sampling rate in dataset B. The GAF image generated by dataset B contains an enormous amount of data. Therefore, the corresponding highlighted lines are thinner in dataset B.

Similarly to dataset A, the data at 1 kHz in dataset B is also tested five times. Table 8 and Figure 22 show that the diagnostic accuracy of the proposed method is 99.91%. The accuracy is 1.5% higher than the SE model, 2.93% higher than the CBAM, and 2.92% higher than the BAM model. The accuracy of the VMD model and EMD model is 71.20% and 93.73%.

Figure 23 shows the t-SNE visualization of diagnostic results of different methods on dataset B. Multiple states in the diagnosis results of the VMD and EMD model overlap. In the BAM model, the state characteristics of OF2 and OF3 are relatively close. Figure 24 shows the confusion matrix of these methods and summarizes the diagnostic results of different methods. VMD model has prediction errors in multiple bearing states, and the EMD model has low diagnostic accuracy in RF3. The results show that GAF-CA-CNN can automatically extract useful features and realize fast, high-precision fault diagnoses.
Table 8. Diagnostics result of different methods on Dataset B.

| Models                  | 1    | 2    | 3    | 4    | 5    | Average | Standard Deviation |
|-------------------------|------|------|------|------|------|---------|--------------------|
| GAF-CA-CNN              | 99.91% | 99.90% | 99.95% | 99.85% | 99.95% | 99.91% | 0.0371            |
| GAF-SE-CNN              | 99.95% | 98.75% | 96.54% | 99.55% | 97.25% | 98.41% | 1.3137            |
| GAF-CBAM-CNN            | 94.10% | 98.40% | 96.20% | 99.30% | 96.90% | 96.98% | 1.8059            |
| GAF-BAM-CNN             | 99.85% | 99.35% | 93.45% | 98.60% | 93.70% | 96.99% | 2.8176            |
| VMD-Gray image-Resnet50 | 67.75% | 75.25% | 68.50% | 70.00% | 74.50% | 71.20% | 3.0959            |
| EMD-Gray image-Resnet50 | 96.67% | 95.33% | 93.00% | 87.33% | 93.73% | 93.73% | 3.4484            |

Figure 22. Comparison results between seven methods on dataset B.

Figure 23. Visualization of different diagnostic results on dataset B, (a) The proposed method, (b) GAF-SE-CNN, (c) GAF-CBAM-CNN, (d) GAF-BAM-CNN, (e) VMD-Gray image-Resnet50, (f) EMD-Gray image-Resnet50.
Figure 24. The confusion matrix of different diagnostic results on dataset B, (a) The proposed method, (b) GAF-SE-CNN, (c) GAF-CBAM-CNN, (d) GAF-BAM-CNN, (e) VMD-Gray image-Resnet50, (f) EMD-Gray image-Resnet50.

5. Conclusions

To improve the fault diagnosis performance of rolling bearing, a lightweight CNN bearing fault intelligent diagnosis model, combining GAF and CA, has been presented. Firstly, the time-series vibration signal is encoded into the GAF images after comparing the performance of different downsampling frequencies. The GAF images preserve the temporal relations and reduce the Gaussian noise to reveal the bearing fault characteristics. Secondly, the lightweight CNN model is realized by deep separable convolution, inverse residual block, and linear bottleneck layer for further feature extraction and classification. Meanwhile, the model introduced the CA to augment the input feature map representations of the GAF images. The results show that almost no additional computational overhead is made under this mechanism. Additionally, verification and evaluation of the proposed method has been processed on the CWRU motor bearing and experimental datasets. It has been demonstrated that GAF-CA-CNN has higher classification performance and less computational overhead than the existing diagnosis methods.

Besides, the DL fault diagnosis model has a robust feature extraction and classification ability, but its performance is affected by data, scale, and quality. By combining time series characteristics and signal processing technology, the model can extract meaningful signal features, which is conducive to the successful application of the DL model in mechanical health detection.

Author Contributions: Conceptualization, J.C. and Q.Z.; methodology, J.C.; software, J.C.; validation, J.C., Q.Z. and S.Z.; formal analysis, Q.Z.; investigation, J.C.; resources, J.C. and Q.Z.; data curation, J.C.; writing—original draft preparation, J.C.; writing—review and editing, J.C. and Q.Z.; visualization, L.P and J.W.; supervision, S.Z.; project administration, Q.Z., S.Z. and L.P.; funding acquisition, S.Z. and L.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 51975347 and Grant No. 51907117) and Shanghai Science and Technology Program (Grant No. 22010501600).
Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Acknowledgments: Heartfelt thanks to Xieqi Chen for English editing to improve the readability of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Glossary

- $X$: Time series
- $x_i$: Normalized value
- $\phi$: The angle of the polar coordinate system
- $r$: The radius of the polar coordinate system
- $t_i$: Timestamp
- $N$: A constant to adjust the radius
- $a_k$: Vectors
- $G$: Gramian angular field matrix
- $X$: Characteristic graph
- $z_{c,h,w}^\rho$: The unidirectional pooling result of the $c$-th channel at height $h$ and weight $w$
- $F_{1,h,w}^1 \times 1$: Convolution function
- $\delta$: Non-linear activation function
- $\sigma$: Sigmoid function
- $H$: The height of the channel
- $W$: The weight of the channel
- $C$: The number of the channel
- $N$: Kernel size
- $n$: The number of the input data
- $f \in \mathbb{R}^{C/r \times (H+W)}$: is a feature map
- $r$: Reduction ratio to control the block size
- $g$: Attention weights
- $V_c$: The result of CA attention

References

1. Udmale, S.S.; Singh, S.K.; Bhirud, S.G. A bearing data analysis based on kurtogram and deep learning sequence models. *Measurement* 2019, 145, 665–677. [CrossRef]
2. Liu, J.; Chen, A.; Zhao, N. An Intelligent Fault Diagnosis Method for Bogie Bearings of Metro Vehicles Based on Weighted Improved D-S Evidence Theory. *Energies* 2018, 11, 232. [CrossRef]
3. Peng, L.; Zheng, S.; Li, P.; Wang, Y.; Zhong, Q. A comprehensive detection system for track geometry using fused vision and inertia. *IEEE Trans. Instrum. Meas.* 2020, 70, 1–15. [CrossRef]
4. Chen, B.; Shen, B.; Chen, F.; Tian, H.; Xiao, W.; Zhang, F.; Zhao, C. Fault diagnosis method based on integration of RSSD and wavelet transform to rolling bearing. *Measurement* 2019, 131, 400–411. [CrossRef]
5. Zair, M.; Rahmoune, C.; Benazzouz, D. Multi-fault diagnosis of rolling bearing using fuzzy entropy of empirical mode decomposition, principal component analysis, and SOM neural network. *Proc. Inst. Mech. Eng. Part. C J. Mech. Eng. Sci.* 2018, 233, 3317–3328. [CrossRef]
6. Ding, J.; Huang, L.; Xiao, D.; Li, X. GMPSO-VMD Algorithm and Its Application to Rolling Bearing Fault Feature Extraction. *Sensors* 2020, 20, 1946. [CrossRef] [PubMed]
7. Li, J.; Wang, H.; Wang, X.; Zhang, Y. Rolling bearing fault diagnosis based on improved adaptive parameterless empirical wavelet transform and sparse denoising. *Measurement* 2019, 152, 107392. [CrossRef]
8. Chen, Q.-Q.; Dai, S.-W.; Dai, H.-D. A Rolling Bearing Fault Diagnosis Method Based on EMD and Quantile Permutation Entropy. *Math. Probl. Eng.* 2019, 2019, 1–8. [CrossRef]
9. Li, X.; Ma, Z.; Kang, D.; Li, X. Fault diagnosis for rolling bearing based on VMD-FRFT. *Measurement* 2020, 155, 107554. [CrossRef]
10. Qiao, Z.; Liu, Y.; Liao, Y. An Improved Method of EWT and Its Application in Rolling Bearings Fault Diagnosis. *Shock Vib.* 2020, 2020, 1–13. [CrossRef]
11. Gunerker, R.S.; Jalan, A.K.; Belgamwar, S.U. Fault diagnosis of rolling element bearing based on artificial neural network. *J. Mech. Sci. Technol.* 2019, 33, 505–511. [CrossRef]
