Multitask Learning Based on Lightweight 1DCNN for Fault Diagnosis of Wheelset Bearings

Zhiliang Liu, *Member, IEEE*, Huan Wang, Junjie Liu, Yong Qin, *Member, IEEE*, and Dandan Peng

***Abstract*— In recent years, deep learning has been proved to be a promising bearing fault diagnosis technology. However, most of the existing methods are based on single-task learning. Fault diagnosis task (FDT) is treated as an independent task, and rich correlation information contained in different tasks is ignored. Therefore, this article explores the possibility of using speed identification task (SIT) and load identification task (LIT) as two auxiliary tasks to improve the performance of the FDT and proposes a multitask one-dimensional convolutional neural network (MT-1DCNN). Specifically, the MT-1DCNN utilizes trunk network to learn shared features required for every task and then processes different tasks through multiple task-specific branches. In this way, the MT-1DCNN can utilize features learned by related tasks to improve the performance of the FDT. The experimental results with wheelset bearing data set show that the multitask learning can make full use of the feature information captured by the SIT and the LIT to improve the fault diagnosis performance of the network, and the MT-1DCNN has a better performance than five excellent networks in accuracy.**

***Index Terms*— Bearing fault diagnosis, convolutional neural network (CNN), multitask learning (MTL), vibration analysis.**

I. Introduction

WHEELSET bearing is a core component of high-speed train (HST) bogie, and its mechanical performance greatly affects the safety and reliability of the HST operation. Therefore, automatic health monitoring for wheelset bearing is of great significance [1]. Due to long-term operation of the HST under time-variant conditions such as speed, load, and operation environment, vibration signals from wheelset bearing are easily interfered. It brings a big challenge for accurate fault diagnosis based on vibration analysis.

The main work of fault diagnosis research is to extract useful information from vibration signals and then use classification methods to obtain the robust diagnosis results. Scholars have proposed various signal processing methods to extract representative features. For example, variational mode decomposition [2], [3], empirical wavelet transform [4], [5], and local mean decomposition [6], [7]. In addition, support vector machine (SVM) [8], [9], k-nearest neighbor [10], [11], and multilayer perceptron [12] are often used as classifiers to predict fault types. For instance, Zheng *et al.* [13] proposed a fault diagnosis method based on multiscale fuzzy entropy and SVM for rolling bearing. Choi *et al.* [14] reduced the diagnostic errors by fusing multiple classifier decisions. However, these methods rely heavily on the domain knowledge of professionals, and they cannot comprehensively extract the complex dynamic features of the signals. Robustness and accuracy of these methods need to be further improved.

As an efficient feature extraction and pattern recognition method, deep learning attracts more and more attention from researchers [15]. In particular, convolutional neural network (CNN) has achieved significant success in fault diagnosis of rotating machinery due to its unique feature learning mechanism through convolution operation [16]–[27]. For example, Liu *et al.* [21] proposed a residual CNN with a multiscale kernel for motor fault diagnosis in nonstationary conditions. Zhang *et al.* [28] proposed a deep CNN with wide first-layer kernels, which can better learn the long-time information of vibration signals. These methods are based on one-dimensional CNN (1DCNN) [17]–[21], [29], which uses the 1DCNN to automatically learn useful information of vibration signals and diagnose the health condition of machinery. In addition, Wen *et al.* [30] transformed vibration signals into two-dimensional (2-D) image and then used 2-D CNN (2DCNN) to learn the useful features. The 2DCNN-based methods usually require converting 1-D signal into 2-D matrix (e.g., time–frequency spectra) [30], [31]. Compared with 2DCNN, 1DCNN can learn the features directly, and the structure is relatively simple, so it is more suitable for bearing fault diagnosis.

The above deep learning networks are all based on single-task learning. Their network parameter optimization is constrained by fault diagnosis task (FDT), and thus, the features learned by the network are only applicable to the diagnosis of mechanical health condition. This approach seems reasonable, but there are implicit shortcomings. Many problems in the real world cannot be decomposed into independent subtasks. Even if it can be decomposed, its subtasks are related to each other and are connected by some sharing factors or sharing features [32]. Therefore, if the real problem is treated as multiple independent single tasks, the rich associated information among these tasks is ignored.

Multitask learning (MTL) [32]–[35] is a machine learning method aimed at solving multiple tasks at the same time. It can use the useful information learned by related tasks to

Manuscript received March 3, 2020; accepted August 5, 2020. Date of publication August 26, 2020; date of current version November 23, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61833002. The Associate Editor coordinating the review process was Zhigang Liu. *(Corresponding authors: Huan Wang; Yong Qin.)*

Zhiliang Liu, Huan Wang, and Junjie Liu are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: wh.huanwang@gmail.com).

Yong Qin is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: yqin@bjtu.edu.cn).

Dandan Peng is with the Department of Mechanical Engineering, KU Leuven, 3001 Leuven, Belgium.

Digital Object Identifier 10.1109/TIM.2020.3017900

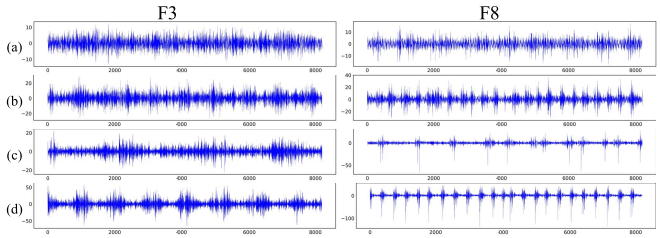

Fig. 1. Vibration signals with different fault categories (F3 and F8) at different speeds and vertical loads. (a) Speed 60 km/h, vertical load 56 KN. (b) Speed 120 km/h, vertical load 56 KN. (c) Speed 60 km/h, vertical load 272 KN. (d) Speed 120 km/h, vertical load 272 KN. (Note: F3 and F8 are described in detail in Section IV).

improve the performance of the network. Caruana *et al.* [32] summarized that the MTL is an approach to inductive transfer that improves the generalization by using domain information contained in the training signals of related tasks as an inductive bias. Simply put, this method can learn the shared features of multiple tasks and allows the features that are specific to a task of these shared features to be used by other tasks, which can effectively promote the results of the main task and other auxiliary tasks. This advantage is not possessed by single-task learning.

There is some research on MTL in the field of fault diagnosis. For example, Guo *et al.* [36] introduced MTL to the field of fault diagnosis and proposed a multitask neural network for processing fault mode task and fault location task simultaneously. Liu *et al.* [37] proposed an MTL method that simultaneously predicts the fault category and remaining useful life. Cao *et al.* [38] used MTL to diagnose the health status of planetary gearbox. These methods proved the effectiveness of the MTL in the FDT but lacks in-depth analysis and interpretation of feature learning mechanism of the MTL. In addition, these works ignore the correlation between the working condition and the health condition.

Vibration response of rotating machinery is related not only to the health condition but also to the working condition, such as speed and load. Fig. 1 shows the vibration responses of the wheelset bearing with two different fault categories at different speeds and loads. The fault category, speed, and load of rotating mechanical system have a great influence on the vibration responses. If we make the network learn these related tasks together and make them share the learned features, the network can have a more comprehensive understanding of the vibration signals, and the learning features also have a better generalization performance.

Therefore, this article introduces the MTL principle into the bearing fault diagnosis and proposes a multitask one-dimensional CNN (MT-1DCNN). The MT-1DCNN aims to enhance the performance of the network by using the two auxiliary tasks: speed identification and load identification. Specifically, the MT-1DCNN processes three tasks at the same time and first learns the shared features among multiple tasks through the trunk network. Subsequently, the MT-1DCNN uses multiple task-specific branches to process these tasks. The input of these branches is the shared features learned by the trunk network. Every task-specific branch can take advantage of the shared features learned by multiple tasks. In this way, the features specific to one task of the shared features can be used by other tasks, so that the network can fully understand the characteristics of the signals and improve the accuracy of each task. In addition, each task has an independent loss function. The overall loss of the MT-1DCNN is obtained by adding up the loss functions of these tasks according to a certain weight. Powered by MTL, the MT-1DCNN can process three tasks simultaneously in a lightweight network structure, and good results can be obtained.

Contributions of this article are summarized as follows.

1) We introduce the MTL principle to wheelset bearing fault diagnosis. The effectiveness of the MTL principle has been demonstrated with implementations based on two deep learning architectures.
2) We propose a lightweight CNN-based network that uses vibration signals to simultaneously deal with three related tasks: FDT, speed identification task (SIT), and load identification task (LIT).
3) We conduct a set of performance comparison with the wheelset bearing data set. In addition, we interpret feature learning mechanism of MTL by using visualization technique.

This article is organized as follows. Section II defines MTL in detail. Section III describes the MT-1DCNN meticulously. Section IV verifies the effectiveness and superiority of the MT-1DCNN with the wheelset bearing data set. Section V discusses four aspects of the MT-1DCNN. Section VI summarizes the whole article.

II. Multitask Learning Concept

Given m learning tasks $\{\Gamma^i\}_{i=1}^m$ where all tasks or a subset of them are related, MTL aims to improve the learning of a model for the Γ^i by using the knowledge contained in all or some of the m tasks [33].

Based on this definition, we focus on supervised learning in MTL since most FDTs fall in this setting. In the setting of supervised learning tasks, usually a task Γ^i is accompanied by a training data set D^i consisting of n^i training samples, i.e., $D^i = \{g_j^i, l_j^i\}_{j=1}^{n^i}$, where $g_j^i \in \mathbb{R}^{d^i}$ is the jth training instance in Γ^i and l_j^i is its corresponding label. Here, we consider a special setting for MTL that the training data D^i for each task is the same. In this setting, the network learns shared features from the same data set that can be used for multiple task processing. This sharing feature can be shared among these tasks, so as to improve the generalization performance of the network. Thus, sharing what is learned by different tasks while tasks are trained in parallel is the central idea of MTL.

III. MT-1DCNN-Based Fault Diagnosis Method

This study is devoted to explore the application of the MTL in improvement of wheelset bearing fault diagnosis. HST working condition (such as speed and load) is closely related to vibration response of wheelset bearing. A fault diagnosis method is expected to integrate all these comprehensive information of wheelset bearings and to improve the diagnosis results. Therefore, this article proposes the MT-1DCNN, which

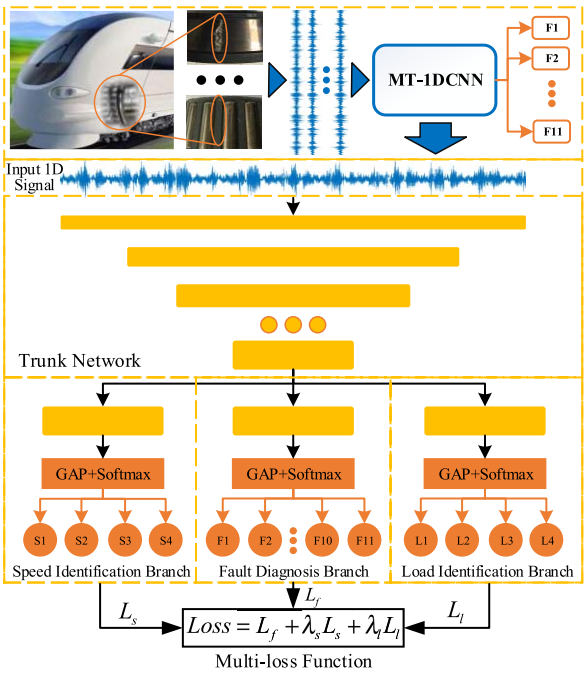

Fig. 2. Overall architecture of the MT-1DCNN.

can learn fault information and working condition information of vibration signals at the same time. Focusing on the FDT, this method introduces the SIT and LIT as two auxiliary tasks to obtain more generalized shared features through multitask collaborative learning. The overall structure of the MT-1DCNN is shown in Fig. 2. The MT-1DCNN mainly consists of three parts: trunk network, task-specific branches, and multiloss function.

A. Trunk Network

In the MT-1DCNN, the trunk network takes the 1-D vibration signal as input and then uses multiple convolutional layers to learn the rich features contained in the raw signal. Different from other CNNs, the trunk network can learn not only the fault-related features but also the features specific to fault-related tasks. That is, the trunk network can learn shared features of multiple related tasks, which contain all the feature information needed to process these tasks. Inspired by Zhang *et al.* [28], we build a lightweight and excellent trunk network, whose structure is described in Fig. 2 and Table I. The trunk network consists of five convolution modules, each of which consists of a convolution layer and a ReLU activation function. The length of the input 1-D signals is 2048 × 1. To capture the long-term features of the signal, the convolution kernel size of the first and the second layers of the trunk network is set to 12 × 1. To reduce the network parameters, we gradually reduce the size of the convolution kernel. In addition, to reduce the complexity of the trunk network, we set the number of channels to 16 in the first convolution layer and then gradually increase to 32. In the network, the stride size of each convolutional layer is set to 2 to achieve the down-sampling,

TABLE I
NETWORK CONFIGURATION OF THE MT-1DCNN ARCHITECTURE

Layer	Layer Type	Kernel Size	Channel	Stride	Padding
1	Conv	12×1	16	2	Yes
2	Conv	12×1	16	2	Yes
3	Conv	9×1	24	2	Yes
4	Conv	9×1	24	2	Yes
5	Conv	6×1	32	2	Yes

Task-specific Branches

Speed Identification			Fault Diagnosis Branch			Load Identification		
Layer Type	Size channel	Stride padding	Layer Type	Size channel	Stride padding	Layer Type	Size channel	Stride padding
Conv	6×1 32	2 Yes	Conv	6×1 32	2 Yes	Conv	6×1 32	2 Yes
Conv	3×1 64	2 Yes	Conv	3×1 64	2 Yes	Conv	3×1 64	2 Yes
Global Average Pooling			Global Average Pooling			Global Average Pooling		
Softmax			Softmax			Softmax		

which can avoid the information loss caused by using the max-pooling. It can be seen that our trunk network can capture the long-term features and short-term features of the input signals with small model complexity and effectively learn the shared features of multiple tasks.

B. Task-Specific Branches

This study aims to use the MTL principle to simultaneously process the FDT, the SIT, and the LIT, so that they can share features with each other and promote the performance of the FDT. Among them, the FDT is to diagnose the health condition of the bearing. The SIT and the LIT are to perceive speed and load of the rotating mechanical system, respectively. To this end, we design three task-specific branches, which are fault diagnosis branch (FDB), speed identification branch (SIB), and load identification branch (LIB). These three branches share the learning features of the trunk network. Therefore, the introduction of the SIB and the LIB enables the trunk network to learn the speed and load information implicitly including in the vibration responses. The FDB can make full use of the rich information of the shared features to accurately distinguish different fault categories.

As shown in Fig. 2 and Table I, suppose that the shared features learned by the trunk network is $M = f^{\mathrm{t}}(X;\theta_{\mathrm{t}}), X \in \mathbb{R}^{T\times 1}$, where X is the input signal of the network, $T = 2048$ is the length of the signal, f^{t} represents the function learned by the trunk network, and θ_{t} is the parameter of f^{t}. The FDB, the SIB, and the LIB take M as input and then use two convolution modules to learn the features (Y^{f}, Y^{s}, and Y^{l}) that are used for specific tasks processing from M. This process can be expressed as follows:

$$Y^{\mathrm{f}}, Y^{\mathrm{s}}, Y^{\mathrm{l}} = f^{\mathrm{f}}(M;\theta_{\mathrm{f}}), f^{\mathrm{s}}(M;\theta_{\mathrm{s}}), f^{\mathrm{l}}(M;\theta_{\mathrm{l}}) \quad (1)$$

where f^{f}, f^{s}, and f^{l} are the feature extraction functions learned by the FDB, SIB, and LIB, respectively, and θ^{f}, θ^{f}, and θ^{l} are the corresponding parameters.

Then, a global average pooling layer (GAP) [39] is used to compress the global information of each channel on Y^{f}, Y^{s},

and Y^{l} into a channel descriptor, so as to get feature vectors z^f, z^s, and z^l. The jth element of z^f is calculated by the following equation:

$$z_j^f = \mathrm{GAP}\left(Y^{\mathrm{f}}\right) = \frac{1}{1 \times W} \sum_{u}^{W} Y_j^{\mathrm{f}}(u), \quad z^f \in \mathbb{R}^{1 \times C} \tag{2}$$

where W and C are the length and number of channels of Y^{f}, respectively.

The GAP can compress the input features into a vector, which greatly reduces the network parameters. Therefore, it can effectively avoid the over fitting problem caused by the full connection layer. In addition, the GAP is more native to the convolution structure by enforcing correspondences between feature channels and categories [39].

Wheelset bearing has many fault types, so the FDT is a multiclass classification problem. In this study, we also consider the SIT and the LIT into multiclass classification problems. The softmax activation function is used for the three tasks. The softmax function maps the input feature vector to a range from 0 to 1 and makes the sum of all elements of the vector equal to 1, so that it is generally used as the classifier to estimate the probability distribution belonging to different classes. We assume that k^{f}, k^{s}, and k^{l} are the number of health condition, speed condition, and load condition, respectively. In this article, $k^{\mathrm{f}} = 11$ and $k^{\mathrm{s}} = k^{\mathrm{l}} = 4$. The softmax function is expressed as follows:

$$Q_j(\hat{z}) = \frac{\exp\left(\hat{z}_j\right)}{\sum_{j=1}^{k} \exp\left(\hat{z}_j\right)}, \quad j = 1, 2, \ldots, k \tag{3}$$

where $\hat{z}_j$ is the jth element of $\hat{z}$, and $\hat{z}$ is the input vector of softmax activation function. $Q_j(\hat{z})$ is the estimated probability distribution of $\hat{z}$ belonging to the jth class.

C. Multiloss Function

When designing a deep neural network, the choice of the loss function is always an important aspect. The mean squared error and mean absolute error often lead to poor performance when used with gradient-based optimization. Some output units that saturate produce very small gradients when combined with these cost functions [40]. Recently, the cross-entropy loss function gets its popularity and has been widely used in classification tasks because of its better performance.

The MT-1DCNN needs to process three different classification tasks simultaneously, so three cross-entropy loss functions are set. They are L_{f} for the FDT, L_{s} for the SIT, and L_{l} for the LIT. These three loss functions are independent of each other. The cross-entropy loss function is mainly used to evaluate the error of the estimated softmax output probability distribution and the target class probability distribution. Suppose p^{f}, p^{s}, and p^{l} are the target distribution of the FDT, the SIT, and the LIT, respectively, and q^{f}, q^{s}, and q^{l} are the estimated distributions of the three tasks. L_{f}, L_{s}, and L_{l} are expressed as follows:

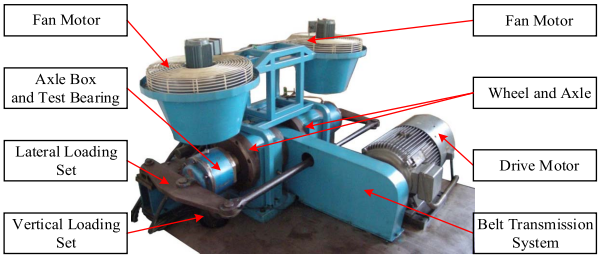

Fig. 3. Wheelset bearing test rig.

$$L_{\mathrm{f}} = -\sum_{j=1}^{k_{\mathrm{f}}} p_j^{\mathrm{f}} \log\left(q_j^{\mathrm{f}}\right); \quad L_{\mathrm{s}} = -\sum_{j=1}^{k_{\mathrm{s}}} p_j^{\mathrm{s}} \log\left(q_j^{\mathrm{s}}\right)$$
$$L_{\mathrm{l}} = -\sum_{j=1}^{k_{\mathrm{l}}} p_j^{\mathrm{l}} \log\left(q_j^{\mathrm{l}}\right). \tag{4}$$

In order to make the three tasks cotrained in the MT-1DCNN and use the features learned by the SIT and the LIT to improve the fault diagnosis performance of the network, adding L_{f}, L_{s}, and L_{l} directly to obtain the final loss function of the network is not optimal. This is because the contribution of different auxiliary tasks to the FDT is inconsistent, and if the auxiliary tasks are given too much weight, the features learned by the network may be more biased to solve the auxiliary tasks. Therefore, we introduce two hyperparameters (i.e., λ_{s} and λ_{l}) to control the weights of the SIT and the LIT, respectively. The total loss of the MT-1DCNN is expressed as follows:

$$\mathrm{Loss} = L_{\mathrm{f}} + \lambda_{\mathrm{s}} L_{\mathrm{s}} + \lambda_{\mathrm{l}} L_{\mathrm{l}}. \tag{5}$$

The accuracy of each task is different, which could result in unbalance problem in the loss function. To address this problem, we introduce two weight coefficients in the loss function to balance accuracies of the three task branches. Selection of the two hyperparameters will be discussed in Section IV-C. Then, we use the stochastic gradient descent method to minimize the loss function in (5). During training, the network simultaneously processes the FDT, the SIT, and the LIT. In other words, the network updates the parameters θ^{t}, θ^{f}, θ^{s}, and θ^{l} at the same time. As training progresses, the trunk network learn more generalized shared features, and task-specific branches can take advantage of the features learned from other tasks to achieve better performance.

IV. Experimental Results and Analysis

In this section, the effectiveness and superiority of MTL and the proposed MT-1DCNN is verified on the wheelset bearing data set.

A. Test Rig for Fault Experiments

As shown in Fig. 3, the wheelset bearing test rig is established to evaluate the effectiveness of the MT-1DCNN. The test rig is designed to simulate a real train operating environment. It is mainly composed of a drive motor,

TABLE II
ELEVEN HEALTH CONDITIONS OF WHEELSET BEARINGS

Location	Fault Description	Class Label
None	Normal	F1
Inner race	Pitting	F2
Rolling element	Pitting	F3
Rolling element	Flaking with a size of 3mm×35mm	F4
Inner race	Flaking with a size of 3mm×45mm	F5
Rolling element	Cracking	F6
Outer race and rolling element	Mixed fault with outer race flaking and rolling element pitting, and the flaking size is 10mm×45mm	F7
Inner race	Flaking with a size of 10mm×45mm	F8
Outer race	Flaking with a size of 10mm×30mm	F9
Rolling element	Flaking with a size of 1mm×1mm	F10
Cage	Cracking	F11

a belt transmission system, and a control system. In addition, we also set up a vertical loading device, a lateral loading device, and two fan motors to simulate the wind resistance and 2-D loads during the real train operation. An axle and its two supporting bearings are assembled to the test rig. The experimental data are collected by acceleration sensors installed in the axle boxes, and the sampling frequency is 5120 Hz.

In this test rig, we collected bearings with naturally generated faults from real operation lines, including a total of 11 health conditions. Table II shows the health condition information of the experimental bearings, which are marked as F1, F2, ..., F11. To simulate the working environment of wheelset bearings on a real train as much as possible, four operation speeds (60, 90, 120, and 150 km/h) and four vertical loads (56, 146, 236, and 272 KN) were set in each healthy condition. Therefore, the SIT has four categories, which are labeled as S1, S2, S3, and S4 and the LIT has four categories, which are labeled as L1, L2, L3, and L4.

B. Experimental Setup

The data are randomly divided into training set, validation set, and test set according to the ratio of 3:1:2, and then sliding segmentation method is used for data augmentation. The sliding segmentation is an efficient vibration signal data augmentation method, which has been used in [16] and [28]. In our experiment, the length of each sample is 2048, and the step size of sliding segmentation is set to 256. Finally, 45 652 training samples, 13 332 validation samples, and 19 700 test samples are obtained. Also, four repeating experiments are carried out for each method.

The MT-1DCNN is implemented by Keras library and Python 3.5. Network training and testing are performed on a workstation with an Ubuntu 16.04 operating system, an Intel Core i9-9900K CPU, and a GTX 2080 GPU with 8G video memory. Each sample accelerates the convergence speed of the network by subtracting the mean and dividing the variance. During the training, the learn rate is 0.0001 and batch size is 96.

Accuracy, recall, and precision are used as evaluation metrics to comprehensively measure the performance of classification methods. They are defined as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \times 100\% \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\% \tag{8}$$

where FP, TP, FN, and TN refer to the number of false-positive samples, true-positive samples, false-negative samples, and true-negative samples, respectively.

To better simulate the strong noise interference of rolling bearings in real operation, we add white Gaussian noise into the raw signals. The definition of SNR is shown as follows:

$$\text{SNR}_{\text{dB}} = 10\log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right) \tag{9}$$

where P_{signal} and P_{noise} are the power of the signal and the noise, respectively.

We compared the MT-1DCNN with the following five excellent networks. First, Wen-2DCNN [30], Guo-2DCNN [31], Wei-1DCNN [28], and Zhang-1DCNN [17] are tested for comparison. In addition the five-layer BPNN is tested for comparison as well. The hidden layers are 1024, 512, 256, 96, and 11. Notably, the training strategy is consistent for every network in all validation.

TABLE III
PERFORMANCE RESULTS FOR DIFFERENT WEIGHTS (SNR = −5 dB)

	$\lambda_s = 1$	$\lambda_s = 0.8$	$\lambda_s = 0.6$	$\lambda_s = 0.4$	$\lambda_s = 0.2$
$\lambda_l = 1$	0.829±0.013	0.829±0.013	0.833±0.012	0.825±0.003	0.812±0.005
$\lambda_l = 0.8$	0.830±0.004	0.829±0.010	0.839±0.003	0.818±0.010	0.815±0.010
$\lambda_l = 0.6$	0.832±0.016	0.818±0.021	0.839±0.009	0.827±0.006	0.821±0.009
$\lambda_l = 0.4$	0.833±0.011	0.841±0.013	**0.847±0.007**	0.818±0.006	0.827±0.010
$\lambda_l = 0.2$	0.825±0.015	0.834±0.015	0.823±0.018	0.826±0.015	0.826±0.011

C. Selecting Weight Coefficients in the Loss Function

In the MT-1DCNN, the two auxiliary tasks have different contributions to the FDT, which may result in the unbalance problem in the loss function. The unbalance problem could be alleviated by tuning the two weight coefficients, whose values are determined by a grid search with the accuracy of the FDB in this article.

In this experiment, λ_s and λ_l are sequentially set to 0.2, 0.4, 0.6, 0.8, and 1 under SNR = −5 dB. In other words, we have conducted 25 different experiments to discuss the influence of weight coefficients on fault diagnosis performance. The accuracy of fault diagnosis under different weight coefficient settings is shown in Table III.

Obviously, different weight settings have an impact on the fault diagnosis performance of the proposed network. This shows that the three tasks interact with each other, and the network can learn their correlation. In addition, it also proves that different auxiliary tasks have different contributions to

TABLE IV
PERFORMANCE RESULTS OF THE FDT (SNR = −5 dB)

Metric \ Method	MT-1DCNN	CNN-FS	CNN-FL	CNN-F
Accuracy	**0.847±0.007**	0.817±0.018	0.800±0.015	0.724±0.028
Recall	**0.837±0.009**	0.805±0.018	0.787±0.017	0.702±0.032
Precision	**0.840±0.007**	0.808±0.020	0.790±0.017	0.712±0.030

the FDT, so it is necessary to set different weight coefficients for different auxiliary tasks. On the other hand, when λ_s and λ_l increase from 0.2 to 1, the fault diagnosis accuracy of the MT-1DCNN increases first and then decreases. This is because a lower weight makes the auxiliary task cannot provide enough contribution, and a higher weight makes the network more inclined to process the auxiliary task. When $\lambda_s = 0.6$ and $\lambda_l = 0.4$, the MT-1DCNN achieves the best fault diagnosis performance. Therefore, in the subsequent experiments, λ_s and λ_l are set to 0.6 and 0.4, respectively.

D. Effectiveness of MTL for Fault Diagnosis Task

In this section, we discuss the impact of the MTL on the FDT. Under SNR = −5 dB, we compare the MT-1DCNN with three network structures. The three networks are CNN-F (only the FDB is included), CNN-FS (including the FDB and the SIB), and CNN-FL (including the FDB and the LIB). The results of the four networks are shown in Table IV.

Compared with the CNN-F, the CNN-FS with the SIB has a more than 9% improvement in the accuracy of the FDT, which shows that the speed information is quite important for the FDT, and it can effectively assist the network in fault diagnosis. In addition, compared with the CNN-F, the CNN-FL with the LIB improves the accuracy of fault diagnosis by 7.5%, which also proves that the load information of rotating mechanical system can promote performance of the FDT. Compared with the LIT, the addition of the SIT improves the fault diagnosis performance of the network. This is because the vibration signals have greater changes at different speeds, so the network can reduce the occurrence of misjudgment with the help of the speed information. Compared with the CNN-FS, the CNN-FL, and the CNN-F, the MT-1DCNN has a considerable improvement in accuracy, recall, and precision. It means that load information and speed information can complement each other and improve the diagnostic performance of the network. The experimental results prove that the MTL can effectively improve the fault diagnosis performance of the network.

Subsequently, we use the *t*-distributed stochastic neighbor embedding (T-SNE) [41] technique to visualize the distribution of the final output of the MT-1DCNN, the CNN-FS, the CNN-FL, and the CNN-F in 2-D embedded space. The visualization results are shown in Fig. 4, where color represents health condition of the wheelset bearing. Coordinates of each point represent its position in the 2-D embedded space, and distance between two points represents their similarity. We use the Fisher score [42] as the metric to quantify the quality of the projection. The Fisher scores are 2.679, 2.298, 1.633, and 1.524 for the four networks. Obviously, the output of

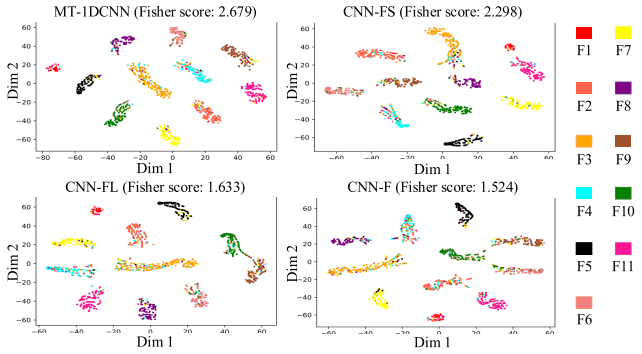

Fig. 4. Group plot visualization for the MT-1DCNN, the CNN-FS, the CNN-FL, and the CNN-F (SNR = −5 dB).

TABLE V
PERFORMANCE RESULTS OF THE SIT (SNR = −5 dB)

Metric \ Method	MT-1DCNN	CNN-FS	CNN-S
Accuracy	**0.887±0.012**	0.878±0.007	0.805±0.004
Recall	**0.888±0.012**	0.879±0.007	0.806±0.005
Precision	**0.887±0.012**	0.880±0.008	0.809±0.004

the network with the MTL ability has a better discrimination ability. In particular, the proposed MT-1DCNN performs at least 14% better than other comparison networks in terms of the Fisher score. It indicates that the MT-1DCNN can effectively improve the feature learning ability of the network.

E. Effectiveness of MTL for SIT

In this section, we discuss the impact of the MTL on the SIT. Under SNR = −5 dB, we set up a comparative experiment of three network structures including the CNN-S (only the SIB is included), the CNN-FS, and the MT-1DCNN. The speed identification results of every network are shown in Table V.

Surprisingly, although the MT-1DCNN did not specifically optimize the SIT, with the addition of the auxiliary tasks, the speed identification performance of the network has been improved. After adding the FDT, the speed identification accuracy of the CNN-FS is 7% higher than that of the CNN-S. The MT-1DCNN is 1% higher than the CNN-FS after the LIT is added. This shows that the MT-1DCNN is not only suitable for the FDT but also effectively improves the performance of related tasks at the same time. This inspires us to explore more methods with MTL in the future work, such as combination of life prediction task and the FDT.

F. Effectiveness of MTL for LIT

In this section, we discuss the impact of the MTL on the LIT. Under SNR = −5 dB, we set up a comparative experiment of three network structures including the CNN-L (only the LIB is included), the CNN-FL, and the MT-1DCNN. The load identification results of each network are shown in Table VI.

Similarly, with the addition of related tasks, the network's load identification performance has been improved. The accuracy of the CNN-FL is 7% higher than that of the CNN-L; the accuracy of the MT-1DCNN is 4% higher than that of the CNN-FL. This again demonstrates the effectiveness of

TABLE VI
PERFORMANCE RESULTS OF THE LIT (SNR = −5 dB)

Metric \ Method	MT-1DCNN	CNN-FL	CNN-L
Accuracy	**0.640±0.012**	0.596±0.011	0.522±0.023
Recall	**0.637±0.012**	0.593±0.011	0.519±0.022
Precision	**0.638±0.013**	0.593±0.009	0.516±0.023

the MTL in dealing with multiple related tasks. However, the performance of the LIT is not particularly good compared with that of the SIT. This is because, due to the existence of standardized operations, the difference of signals under different loads is severely eliminated. Second, this is a result obtained under SNR = −5 dB, and strong noise interferes with the judgment of the network. Third, the goal of the MT-1DCNN is to improve the performance of the FDT, so there is no special optimization for auxiliary tasks.

G. Comparison With Excellent Methods

In this section, we compare the diagnostic performance of the MT-1DCNN and five excellent networks under four SNR scenarios (10, 5, 0, and −5 dB), which are used to simulate the working condition of the wheelset bearing under different noise levels. The fault diagnosis results of every network are shown in Table VII.

Obviously, accuracy, recall, and precision of the MT-1DCNN are better than that of other comparison methods under the four SNR scenarios. In particular, when the noise is strong, the MT-1DCNN is considerably better than other comparison methods in terms of diagnostic accuracy. For example, when SNR = −5 dB, the fault diagnosis accuracy of the MT-1DCNN is more than 13% higher than that of the Wen-2DCNN, which has the best performance among the comparison methods. It means that the MT-1DCNN has a relatively strong antinoise ability even though without any additional denoising preprocessing. With increasing of SNR, the fault diagnosis performance of the network increases. Under SNR = 0 dB, which means the power of noise is about one time of the original signal power, the MT-1DCNN still obtains 96.3% in accuracy of fault diagnosis. Under this noise, the Wen-2DCNN obtained the best results in the comparison methods, but the accuracy is only 91.7%. Under SNR = 10 dB, accuracy of the MT-1DCNN can achieve more than 99.4%. It indicates that even in the noisy background environment, the MT-1DCNN can still achieve a great diagnostic performance and has certain practical application potential. Moreover, the standard deviation of the MT-1DCNN is smaller than other methods in most cases, which shows a good stability.

In order to analyze the diagnosis results of each category and to understand its precision and recall in detail, confusion matrices of the proposed MT-1DCNN under SNR = 10 dB and SNR = −5 dB are provided in Figs. 5 and 6, respectively, where row represents the predicted label and column represents the true label; the diagonal is the number of accurate diagnoses for every categories; the bottom row shows the precision of every categories; and the rightmost column represents the number of testing samples of every

True Label \ Predicted Label	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	Recall	Test
F1	3402	0	6	0	0	0	0	0	0	0	0	**0.9982**	3408
F2	4	1551	0	7	0	5	0	0	0	0	1	**0.9892**	1568
F3	0	4	2632	7	0	0	0	5	0	0	0	**0.9940**	2648
F4	0	2	29	1456	0	0	0	0	0	0	1	**0.9785**	1488
F5	0	0	0	0	1372	0	0	0	0	0	0	**1.0000**	1372
F6	0	1	0	0	0	1562	0	0	2	3	0	**0.9962**	1568
F7	0	0	13	0	6	0	1549	0	0	0	0	**0.9879**	1568
F8	0	0	10	2	0	0	0	1452	0	0	0	**0.9918**	1464
F9	0	6	0	0	0	1	0	1	1560	0	0	**0.9949**	1568
F10	0	0	0	0	0	13	0	0	0	1483	0	**0.9913**	1496
F11	0	4	0	0	0	0	0	0	0	0	1548	**0.9974**	1552
PRE	**0.9988**	**0.9892**	**0.9784**	**0.9891**	**0.9956**	**0.9880**	**1.0000**	**0.9959**	**0.9987**	**0.9980**	**0.9987**	—	19700

Fig. 5. Confusion matrix of the proposed MT-1DCNN (SNR = 10 dB).

True Label \ Predicted Label	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	Recall	Test
F1	3265	22	17	10	0	1	0	6	6	7	74	**0.9580**	3408
F2	56	1244	24	37	0	59	4	1	29	48	66	**0.7934**	1568
F3	5	15	2184	195	14	21	61	74	27	34	18	**0.8248**	2648
F4	4	35	164	1036	3	58	1	33	39	63	52	**0.6962**	1488
F5	0	0	15	0	1331	0	7	15	0	4	0	**0.9701**	1372
F6	0	58	16	34	0	1260	0	1	32	146	21	**0.8036**	1568
F7	0	1	65	9	49	9	1322	47	52	13	1	**0.8431**	1568
F8	0	5	95	74	36	12	17	1190	29	6	0	**0.8128**	1464
F9	0	46	35	40	4	20	41	14	1330	17	21	**0.8482**	1568
F10	1	39	64	29	0	115	11	8	46	1168	15	**0.7807**	1496
F11	86	44	33	48	0	32	0	2	26	29	1252	**0.8067**	1552
PRE	**0.9555**	**0.8244**	**0.8053**	**0.6852**	**0.9262**	**0.7940**	**0.9030**	**0.8555**	**0.8230**	**0.7609**	**0.8237**	—	19700

Fig. 6. Confusion matrix of the proposed MT-1DCNN (SNR = −5 dB).

categories. It is seen that the main error comes from the wrong classification between fault modes. According to their recall values, F4 is the most confusing fault type for the MT-1DCNN. The major of wrongly classified F4 samples are classified into F3. In addition, under SNR = 10 dB, the recall and the precision of the network in most fault modes are close to 100%, which indicates that the MT-1DCNN has good diagnostic performance when the noise is weak. Even under SNR = −5 dB, precision of the MT-1DCNN for the normal category can still reach 95.55%, which shows that the network can still distinguish normal samples and fault samples well under strong noise.

H. Computational Burden of the Networks

Computational burden of the networks is an important metric to measure the performance of bearing fault diagnosis methods. So, this section quantitatively calculates the number of parameters and test time for each method. Table VIII summarizes the total number of parameters and one batch size (96 samples) test time for every network. It is seen that the MT-1DCNN has a very lightweight structure, which obtains better performance than other networks (such as the Wen-2DCNN) by using a small number of parameters, which proves that the MT-1DCNN has a higher parameter utilization rate. However, the MT-1DCNN needs more test time. This is not surprising, since the MT-1DCNN has to process three different tasks simultaneously, which obviously leads to more test time. It is worth pointing out that the test time of the MT-1DCNN on 96 samples is 1.204 s, which is acceptable in engineering practice.

V. DISCUSSIONS ABOUT THE MT-1DCNN

A. Understanding Feature Learning Mechanism of MTL

To understand the feature learning mechanism of the MT-1DCNN, we use the T-SNE [39] technology to visualize

TABLE VII
PERFORMANCE RESULTS OF THE MT-1DCNN AND THE FIVE COMPARISON NETWORKS UNDER THE FOUR SNR SCENARIOS

SNR	Metric \ Method	MT-1DCNN	Wei-1DCNN	Wen-2DCNN	Zhang-1DCNN	Guo-2DCNN	BPNN
10 dB	Accuracy	**0.994±0.001**	0.981±0.005	0.991±0.001	0.988±0.003	0.894±0.011	0.670±0.003
	Recall	**0.993±0.001**	0.980±0.005	0.990±0.001	0.988±0.002	0.886±0.012	0.659±0.007
	Precision	**0.993±0.001**	0.980±0.005	0.990±0.001	0.987±0.002	0.887±0.011	0.661±0.006
5 dB	Accuracy	**0.985±0.001**	0.966±0.006	0.980±0.002	0.964±0.006	0.840±0.010	0.621±0.004
	Recall	**0.984±0.001**	0.964±0.007	0.978±0.002	0.963±0.010	0.826±0.010	0.602±0.004
	Precision	**0.984±0.001**	0.964±0.007	0.979±0.002	0.963±0.006	0.828±0.012	0.608±0.003
0 dB	Accuracy	**0.963±0.004**	0.894±0.009	0.917±0.004	0.866±0.015	0.754±0.010	0.528±0.001
	Recall	**0.962±0.004**	0.889±0.009	0.910±0.005	0.853±0.017	0.736±0.011	0.499±0.002
	Precision	**0.960±0.004**	0.889±0.010	0.910±0.004	0.864±0.017	0.737±0.010	0.508±0.004
−5 dB	Accuracy	**0.847±0.007**	0.666±0.008	0.712±0.005	0.656±0.024	0.546±0.006	0.363±0.003
	Recall	**0.837±0.009**	0.640±0.008	0.686±0.007	0.636±0.019	0.510±0.007	0.317±0.003
	Precision	**0.840±0.007**	0.648±0.015	0.693±0.007	0.651±0.021	0.517±0.008	0.330±0.003

TABLE VIII
NUMBER OF PARAMETERS AND ONE BATCH SIZE (96 SAMPLES) TEST TIME FOR THE MT-1DCNN AND THE FIVE COMPARISON NETWORKS

Metric \ Method	MT-1DCNN	Wei-1DCNN	Wen-2DCNN	Zhang-1DCNN	Guo-2DCNN	BPNN
Number of Parameter	5.5×10^4	5.4×10^4	5.9×10^5	1.3×10^5	1.5×10^4	2.7×10^6
Test Time/s	1.204	0.563	0.505	0.569	0.501	0.497

the distribution of features in different layers of the network in the 2-D space. Fig. 7 shows the visualization of the shallow features, the shared features of the trunk network, the features of every task-specific branch, and the final output.

It is seen that the shallow features learned by the trunk network do not contain specific task information. As the network deepens, under the constraints of related tasks, the trunk network learns the domain-specific information required for multiple related tasks. So, the shared features [Fig. 7(A2), (B2), and (C2)] learnt by the trunk network contain information that can be used for related tasks. Although the discrimination of the shared features is not particularly obvious, relevant auxiliary tasks can provide additional supervision information and make the features learned by the network have a better generalization ability. This also brings another benefit, that is, the local minima of different tasks are in different positions in MTL networks. Through joint learning, it can help the network to escape the local minima. In a single-task learning network, gradient backpropagation tends to fall into local minima. Then, these shared features are sent to the task-specific branches, from which the features that can be used for specific tasks are selected. In this way, the network allows the features that are dedicated to one task of the shared features to be used by other tasks, and such features are often not easy to learn in a single-task learning network. According to Tables IV–VI, we observe consistent results, that is, the classification results of the current task are improved after introducing auxiliary tasks. This shows that certain features that are dedicated to one task are, indeed, useful for its related tasks. Finally, the final results are obtained by classifiers of different tasks. The proposed MT-1DCNN can learn the shared features for multiple related tasks and can also process each task separately. This preserves the independence of each task and allows them to connect and promote each other. It is worth noting that multiple tasks handled by the MT-1DCNN should be related in some extent, which is assumption behind the proposed method.

TABLE IX
PERFORMANCE RESULTS OF THE MT-LSTM AND LSTM UNDER THE FOUR SNR SCENARIOS

Method \ SNR	10 dB	5 dB	0 dB	−5 dB
LSTM	0.967±0.002	0.954±0.003	0.914±0.005	0.739±0.012
MT-LSTM	**0.987±0.001**	**0.981±0.002**	**0.948±0.005**	**0.818±0.010**

B. Combining the MTL Principle With Other Architecture

This section explores the applicability of the proposed multitask principle with other deep learning architectures. We construct two network architectures, namely, long short-term memory (LSTM) and multitask LSTM (MT-LSTM). We first design the LSTM with two-layer LSTM cell, where the length of time steps is 64 and the dimension of input size is 32. Then, we replace the trunk network in the MT-1DCNN with LSTM, keep other network structure the same as in the MT-1DCNN, and finally, construct MT-LSTM. The fault diagnosis accuracy of the two networks under the four SNR scenarios (10, 5, 0, and −5 dB) is shown in Table IX. The experimental results show that the MTL principle successfully combine with LSTM to improve its fault diagnosis performance.

C. Experiments on the CWRU Bearing Data Set

In this section, the proposed method is tested on the bearing data set of Case Western Reserve University (CWRU). The CWRU data set is a public data set. In this data set, a total of four load conditions are set, which are 0, 1, 2, and 3 hp. The data used in this experiment come from the drive end of the test bench motor and contains four different health

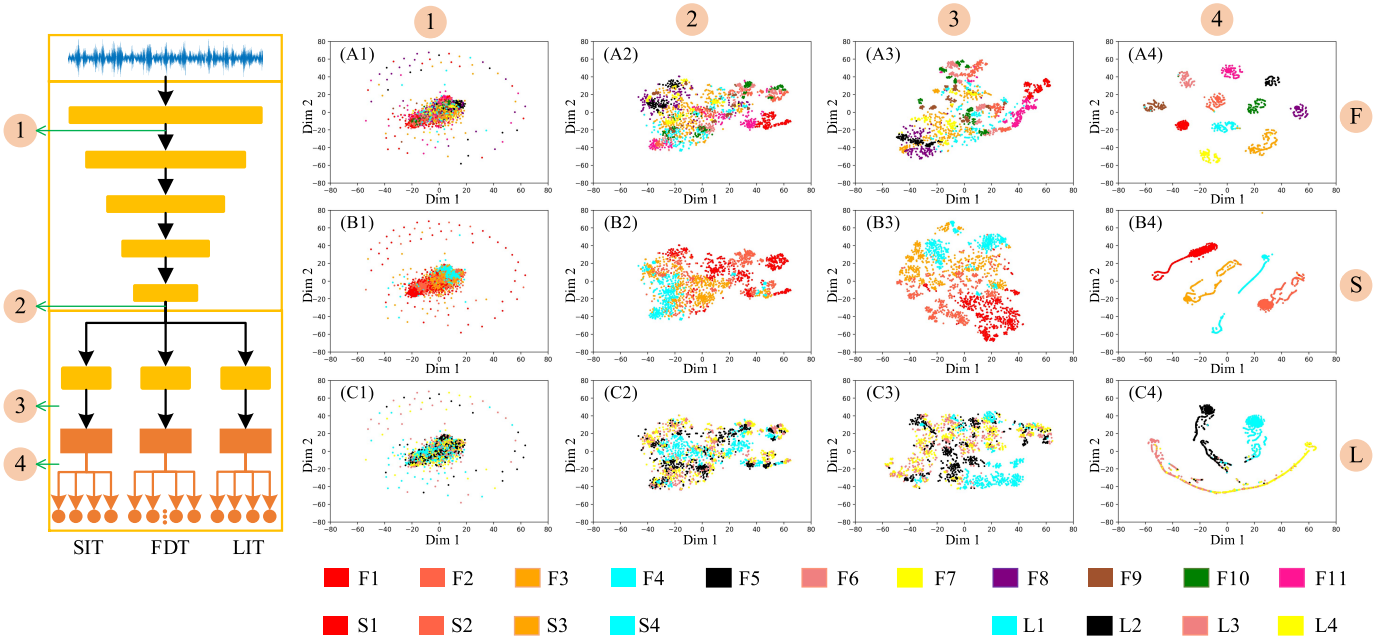

Fig. 7. Group plot visualization of 2-D features in different layers for the MT-1DCNN under SNR = 10 dB, which are visualized by (A1)–(A4) fault categories, (B1)–(B4) speed categories, and (C1–C4) load categories.

TABLE X

PERFORMANCE RESULTS OF THE SIX NETWORKS IN THE CWRU BEARING DATA SET (SNR = −5 dB)

Method \ Metric	Accuracy	Recall	Precision
MT-1DCNN	**0.938±0.007**	**0.938±0.007**	**0.939±0.007**
Wei-1DCNN	0.837±0.015	0.837±0.015	0.836±0.016
Wen-2DCNN	0.863±0.009	0.863±0.009	0.864±0.008
Zhang-1DCNN	0.880±0.002	0.880±0.002	0.889±0.005
Guo-2DCNN	0.621±0.021	0.620±0.022	0.625±0.018
BPNN	0.317±0.002	0.317±0.003	0.326±0.003

states, namely, healthy state, inner ring fault, outer ring fault, and ball fault. All three types of faults are produced by electrodischarge machining. Their diameters are 7, 14, and 21 mils. We treat different degrees of failure as an independent bearing health condition. Therefore, this data set contains ten health conditions and four load conditions. After data augmentation, 9320 training samples, 4200 validation samples, and 4200 test samples are obtained. The fault diagnosis results of six networks under SNR = −5 dB are shown in Table X. It is worth noting that the MT-1DCNN contains only the FDB and LIB. The experimental results show that the MT-1DCNN obtains 93.8% fault diagnosis accuracy, which is an increase of 5.8% compared to Zhang-1DCNN. This indicates that the MT-1DCNN also performs well on the CWRU bearing data set.

D. Novelties of the MT-1DCNN

Condition of the wheelset bearing is restricted by many factors. As shown in Fig. 1, the vibration response of the wheelset bearing is related not only to its health condition but also to its working conditions, such as speed and load. Therefore, if a fault diagnosis model can obtain both health condition information and working condition information, the model can understand the bearing more comprehensively and thereby can make a more accurate decision. However, based on our literature review [36]–[38], we find that current deep learning-based methods ignore association between working condition and health condition. Therefore, this article explores possibility of enabling the network to simultaneously handle working condition identification tasks and the FDT. We prove that MTL can make the network effectively use and share the features learned by different tasks, so as to improve the fault diagnosis performance. The novel MT-1DCNN is proposed, which has achieved very competitive performance on the wheelset bearing data set compared to five peer networks.

The MT-1DCNN first tries to use the multitask structure to benefit the working condition information to improve the fault diagnosis performance of the network. In this way, our method provides a new solution for FDT, and a general and scalable architecture. The MT-1DCNN can be easily expanded if there are other available working condition information (such as temperature). In addition, some task branches can also be reduced. For example, on the CWRU data set, the MT-1DCNN removes the SIB and it still achieves good performance. However, the introduction of more tasks will definitely bring more parameters and calculation cast, and it also challenges the trunk networks' feature learning ability. Therefore, we should propose a lightweight trunk network with stronger feature learning abilities. In this article, inspired by Zhang *et al.* [28], we apply the wide convolution kernel to the entire network to make the network have a more powerful ability to learn long-term correlation features. In addition, we gradually reduce the size of the convolution kernels and construct a shallow network architecture to reduce the network parameters. As shown in Tables VII and VIII, the MT-1DCNN has the same amount of parameters as Wei-1DCNN [28], but it can handle multiple tasks and has a better diagnostic

performance. This indicates that our network improves the feature learning ability while maintaining fewer parameters. Finally, by visualizing the internal feature learning situation of the MT-1DCNN, we discussed the mechanism of the MTL, which also made efforts for interpreting CNN in fault diagnosis field.

E. Overfitting Problem

Multiple tasks are related but not the same. This diversity can reduce the overfitting problem when learning parameters shared in the trunk network. The more tasks we are learning simultaneously, the more our model has to find a representation that captures all the tasks and the less is our chance of overfitting on our original task [43].

However, overfitting is a common problem for supervised learning-based networks, especially for a complex model. During the training process, the network may lead to an overfitted response in which the methods prioritize signal differentiation instead patterns' characterization. In this regard, the features considered by the methods could be specific of the test set (e.g., noise level or electrical interferences) and not common patterns, useful to be applied in the application of the trained methodology in other similar systems. The following two ideas can improve the generalization performance of the network. First, unsupervised learning. Before supervised learning, try to use unsupervised learning to characterize the patterns and then use supervised learning to classify, the overfitting problem can be alleviated. Second, transfer learning. Transferring the patterns' characterization learned by the network on other large data sets to the current task to alleviate the network's overfitted response to the current data set.

VI. Conclusion

This article proposes the end-to-end MT-1DCNN for fault diagnosis of wheelset bearing. It introduces the MTL principle into bearing fault diagnosis and explores the influence of speed information and load information on the FDT. The MT-1DCNN acquires the shared features between these tasks by simultaneously processing the FDT, the SIT, and the LIT. Then, the network can obtain speed information and load information that can assist the classifier for fault diagnosis from the shared features. Therefore, the MT-1DCNN has a more comprehensive feature learning mechanism, which can achieve better performance with a lightweight network structure. The MT-1DCNN establishes a multitask network framework for bearing fault diagnosis, which can not only use speed and load tasks as auxiliary tasks but also further use wind speed, temperature, and other tasks that are related to the FDT. The experimental results show that the MT-1DCNN has considerable advantages over the five peer networks in accuracy, precision, and recall. We also prove that the MTL principle can simultaneously improve the performance of the FDT, the SIT, and the LIT. Moreover, possibility of combining the MTL principle with another network architecture is preliminarily validated.

References

[1] H. Cao, F. Fan, K. Zhou, and Z. He, "Wheel-bearing fault diagnosis of trains using empirical wavelet transform," *Measurement*, vol. 82, pp. 439–449, Mar. 2016.

[2] Z. Li, J. Chen, Y. Zi, and J. Pan, "Independence-oriented VMD to identify fault feature for wheel set bearing fault diagnosis of high speed locomotive," *Mech. Syst. Signal Process.*, vol. 85, pp. 512–529, Feb. 2017.

[3] X.-B. Wang, Z.-X. Yang, and X.-A. Yan, "Novel particle swarm optimization-based variational mode decomposition method for the fault diagnosis of complex rotating machinery," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 68–79, Feb. 2018.

[4] X. Zhang, J. Wang, Z. Liu, and J. Wang, "Weak feature enhancement in machinery fault diagnosis using empirical wavelet transform and an improved adaptive bistable stochastic resonance," *ISA Trans.*, vol. 84, pp. 283–295, Jan. 2019.

[5] Y. Kong, T. Wang, and F. Chu, "Meshing frequency modulation assisted empirical wavelet transform for fault diagnosis of wind turbine planetary ring gear," *Renew. Energy*, vol. 132, pp. 1373–1388, Mar. 2019.

[6] Z. Liu, Y. Jin, M. J. Zuo, and Z. Feng, "Time-frequency representation based on robust local mean decomposition for multicomponent AM-FM signal analysis," *Mech. Syst. Signal Process.*, vol. 95, pp. 468–487, Oct. 2017.

[7] Z. Liu, M. J. Zuo, Y. Jin, D. Pan, and Y. Qin, "Improved local mean decomposition for modulation information mining and its application to machinery fault diagnosis," *J. Sound Vib.*, vol. 397, pp. 266–281, Jun. 2017.

[8] M. Kang, J. Kim, J.-M. Kim, A. C. C. Tan, E. Y. Kim, and B.-K. Choi, "Reliable fault diagnosis for low-speed bearings using individually trained support vector machines with kernel discriminative feature analysis," *IEEE Trans. Power Electron.*, vol. 30, no. 5, pp. 2786–2797, May 2015.

[9] L. Ren, W. Lv, S. Jiang, and Y. Xiao, "Fault diagnosis using a joint model based on sparse representation and SVM," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 10, pp. 2313–2320, Oct. 2016.

[10] P. Baraldi, F. Cannarile, F. Di Maio, and E. Zio, "Hierarchical k-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions," *Eng. Appl. Artif. Intell.*, vol. 56, pp. 1–13, Nov. 2016.

[11] D. H. Pandya, S. H. Upadhyay, and S. P. Harsha, "Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4137–4145, Aug. 2013.

[12] V. N. Ghate and S. V. Dudul, "Optimal MLP neural network classifier for fault detection of three phase induction motor," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3468–3481, Apr. 2010.

[13] J. Zheng, H. Pan, and J. Cheng, "Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines," *Mech. Syst. Signal Process.*, vol. 85, pp. 746–759, Feb. 2017.

[14] K. Choi *et al.*, "Novel classifier fusion approaches for fault diagnosis in automotive systems," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 3, pp. 602–611, Mar. 2009.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] D. Peng, Z. Liu, H. Wang, Y. Qin, and L. Jia, "A novel deeper one-dimensional CNN with residual learning for fault diagnosis of wheelset bearings in high-speed trains," *IEEE Access*, vol. 7, pp. 10278–10293, 2019.

[17] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," *ISA Trans.*, vol. 95, pp. 295–305, Dec. 2019.

[18] H. Wang, Z. Liu, D. Peng, and Y. Qin, "Understanding and learning discriminant features based on multiattention 1DCNN for wheelset bearing fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 5735–5745, Sep. 2020.

[19] J. Pan, Y. Zi, J. Chen, Z. Zhou, and B. Wang, "LiftingNet: A novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 4973–4982, Jun. 2018.

[20] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Deep coupled dense convolutional network with complementary data for intelligent fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9858–9867, Dec. 2019.

[21] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3797–3806, Jun. 2020.

[22] G. Xu, M. Liu, Z. Jiang, W. Shen, and C. Huang, "Online fault diagnosis method based on transfer convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 509–520, Feb. 2020.

[23] I.-H. Kao, W.-J. Wang, Y.-H. Lai, and J.-W. Perng, "Analysis of permanent magnet synchronous motor fault diagnosis based on learning," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 2, pp. 310–324, Feb. 2019.

[24] L. Wen, X. Li, and L. Gao, "A new two-level hierarchical diagnosis network based on convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 330–338, Feb. 2020.
[25] R. Huang, J. Li, W. Li, and L. Cui, "Deep ensemble capsule network for intelligent compound fault diagnosis using multisensory data," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2304–2314, May 2020.
[26] X. Ding and Q. He, "Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 8, pp. 1926–1935, Aug. 2017.
[27] D. Peng, H. Wang, Z. Liu, W. Zhang, M. J. Zuo, and J. Chen, "Multibranch and multiscale CNN for fault diagnosis of wheelset bearings under strong noise and variable load condition," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4949–4960, Jul. 2020.
[28] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, Feb. 2017.
[29] L. Su, L. Ma, N. Qin, D. Huang, and A. H. Kemp, "Fault diagnosis of high-speed train Bogie by residual-squeeze net," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 3856–3863, Mar. 2019.
[30] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.
[31] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Measurement*, vol. 93, pp. 490–502, Nov. 2016.
[32] R. Caruana, "Multitask Learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997.
[33] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, Jan. 2018.
[34] K.-H. Thung and C.-Y. Wee, "A brief review on multi-task learning," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29705–29725, Nov. 2018.
[35] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe, "A multi-task learning framework for head pose estimation under target motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1070–1083, Jun. 2016.
[36] S. Guo, B. Zhang, T. Yang, D. Lyu, and W. Gao, "Multitask convolutional neural network with information fusion for bearing fault diagnosis and localization," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 8005–8015, Sep. 2020.
[37] R. Liu, B. Yang, and A. G. Hauptmann, "Simultaneous bearing fault recognition and remaining useful life prediction using joint-loss convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 87–96, Jan. 2020.
[38] X. Cao, B. Chen, and N. Zeng, "A deep domain adaption model with multi-task networks for planetary gearbox fault diagnosis," *Neurocomputing*, vol. 409, pp. 173–190, Oct. 2020.
[39] L. Min, C. Qiang, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400
[40] I. Goodfellow, Y. Bengio, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[41] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 2605, pp. 2579–2605, Nov. 2008.
[42] Z. Wang and Z. Qian, "Effects of concentration and size of silt particles on the performance of a double-suction centrifugal pump," *Energy*, vol. 123, pp. 36–46, Mar. 2017.
[43] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*. [Online]. Available: http://arxiv.org/abs/1706.05098

Zhiliang Liu (Member, IEEE) was born in Rizhao, Shandong, China, in 1984. He received the Ph.D. degree from the School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2013.

From 2009 to 2011, he was a Visiting Scholar with the University of Alberta, Edmonton, AB, Canada, for two years. From 2013 to 2015, he was an Assistant Professor with the School of Mechanical and Electrical Engineering, UESTC. Since 2015, he has been an Associate Professor with the School of Mechanical and Electrical Engineering. He has authored over 70 articles including 20+ SCI-indexed journal articles. He currently holds 10+ research grants from the National Natural Science Foundation of China, Open Grants of National Key Laboratory, and China Postdoctoral Science Foundation. His research interests include fault diagnosis and prognostics of rotating machinery by using advanced signal processing and data mining methods.

Huan Wang was born in Hunan, China. He received the B.S. degree from the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2016, where he is currently pursuing the M.S. degree with the School of Mechanical and Electrical Engineering.

His research interests include mechanical fault diagnosis, image recognition, deep learning, and machine learning.

Junjie Liu was born in Chongqing, China, in 1994. He received the B.S. degree in mechanical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, where he is currently pursuing the M.S. degree in mechanical engineering.

His research interests include transfer learning, equipment reliability, fault diagnosis, and health management.

Yong Qin (Member, IEEE) received the B.Sc. and M.Sc. degrees in transportation automation and control engineering from Shanghai Railway University, Shanghai, China, in 1993 and 1996, respectively, and the Ph.D. degree in information engineering and control from the China Academy of Railway Sciences, Beijing, China, in 1999.

He is currently a Professor and a Vice Dean of the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing. He has authored or co-authored over 100 publication articles (SCI/EI), one ESI highly cited article, and five books and also has 23 patents granted including two U.S. patents. His research area mainly focused on prognostics and health management for railway transportation system, transportation network safety and reliability, and rail operation planning and optimization.

Dr. Qin is a member of the IEEE ITS and RS and a Senior Member of IET. He won 11 Science and Technology Progress Award of Ministry.

Dandan Peng was born in Shanxi, China, in 1992. She received the B.S. and M.S. degrees from the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2016 and 2019, respectively. She is currently pursuing the Ph.D. degree in mechanical engineering with KU Leuven, Leuven, Belgium.

Her research interests include Hilbert–Huang transform, convolutional neural network, machinery condition monitoring, and fault diagnosis.