Learn Generalization Feature via Convolutional Neural Network: A Fault Diagnosis Scheme Toward Unseen Operating Conditions

YUANTAO YANG, JIANCHENG YIN, HUAILIANG ZHENG, YUQING LI, MINQIANG XU, AND YUSHU CHEN
Deep Space Exploration Research Center, Harbin Institute of Technology, Harbin 150001, China
Corresponding author: Minqiang Xu (xumq@hit.edu.cn)

ABSTRACT In recent years, Convolutional neural networks (CNNs) have achieved start-of-art performance in the fault diagnosis field. If there is no available information on the unseen operating conditions, the model trained on the seen operating condition cannot perform well. One of the feasible strategies is to enhance the generalization ability of the network on various seen operating conditions. We introduce the center loss to the traditional CNN and build an end-to-end fault diagnosis framework (called CNN-C). By minimizing the intra-class variations, center loss cluster the learned features across various seen operating conditions. With the joint supervision of the center loss and the softmax loss, the learned features of the same class could minimize the domain difference across various seen operating conditions while the features of different classes are separable. The generalization ability of network is improved on unseen operating conditions. Compared with the shallow methods and traditional CNN, the proposed method is promising to deal with the fault diagnosis tasks of the bearing and gearbox.

INDEX TERMS Convolutional neural network, center loss, unseen operating condition, fault diagnosis, feature generalization.

I. INTRODUCTION
Fault diagnosis of rotating machinery is curial to reduce maintenance costs and improve the safety and reliability of the systems [1]–[3]. In recent years, Convolutional neural networks (CNNs) have been widely used and shown to be effective in the fault diagnosis field [4]–[6]. As a deep learning algorithm, CNN could automatically learn the nonlinear map between the input original signals and the output class labels, which frees the human labor from the signal processing skills and expert domain knowledge. Janssens et al. [7] used the CNN model to diagnose bearing faults with the input of the frequency spectrum of two accelerometers. Yang et al. [8] combined the hierarchical symbol analysis with CNN and verified the model on a centrifugal pump dataset and a bearing dataset. Jia et al. [9] designed the weight normalization strategy and weighted softmax loss to deal with the imbalanced classification problem. Haidong et al. [10], He et al. [11] combined deep neural networks with wavelet analysis methods to diagnosis the gearbox faults and the early fault of bearing.

Generally, in the fault diagnosis field of the rotating machinery system, the operating conditions are always complex and various such as the shaft speed and workload. Based on the plenty of the vibration signals are collected by the sensors, signal processing skills and deep learning algorithms are utilized to extract the fault characteristics of different health conditions. These features are extracted from the seen operating conditions, which means that both the training samples and testing samples are followed the same distribution. However, most operating conditions are unseen to the fault diagnosis model. It is necessary to discuss the performance of these well-trained models on the unseen operating conditions. In this situation, there is an immediate need is to further improve the generalization ability of the fault diagnosis framework.

However, the above research works only focused on learning discriminative features and improving the classification accuracies but ignored the generalization properties of the learned features. As a result, in the feature spaces learned
by deep neural networks, although the samples of different classes are distinguishable, the samples of the same class may separate into several different clusters, and this phenomenon is more obvious under the data from multiple operating conditions. In other words, although these features are discriminative enough to distinguish from each other and achieve excellent classification accuracy, the generalization ability of the features is poor.

In recently, plenty of transfer learning methods are proposed to deal with the cross-domain fault diagnosis task. The core idea of transfer learning methods is learning a robust model for the testing domain of which the data distribution is related but different from the one of the training domain [12]. Most methods are usually based on the assumption that the information on the testing domain is known. However, in many other scenarios, there is no information on the target domain in the training progress. The fault diagnosis task on the unseen operating conditions is a typical application scenario.

According to the review [13], several deep networks with specific structure and training strategies are applied to reduce the gap between the operating conditions. Zhang et al. [14] presented the CNN model with dropout in the first layer, small batch size, and ensemble learning strategy. CWRU bearing dataset is used to verify the generalization of the model. Han et al. [15]–[18] utilized the spatiotemporal pattern network to explore the relationship among the multiple sensors and diagnose the unseen operating conditions of wind turbine data. Chen et al. [19] employed the deep inception net with atrous convolution to extract common features between the artificial data and the natural data in the Paderborn bearing dataset. Zhu et al. [20] studied CNN based on the capsule network to improve the generalization and applied on CWRU bearing dataset and Paderborn bearing dataset. Peng et al. [21] developed a noise deep convolution neural network to identify the fault of RC reducer under different operating conditions. Wen et al. [22] designed snapshot ensemble CNN to improve the generalization ability and applied on CWRU bearing dataset, MFPT bearing dataset and self-priming centrifugal pump dataset. Wang et al. [12], [23] investigated the probabilistic transfer factor analysis for diagnosis tasks across various operating conditions and applied on a gearbox dataset. Besides, Wei et al. [24] addressed the generalization problem of the rotating speed fluctuation by a rotating speed normalization approach and tested on a case study of rotor crack diagnosis.

In this paper, we focus on the scenario that the fault diagnosis tasks on the unseen operating conditions. First, the training samples are collected from various seen operating conditions while the testing samples are collected from unseen operating conditions. Second, there is no information about the unseen operating condition in the training progress, such as load, shaft speed, normal samples, and fault samples. One of the possible strategies is to improve the domain generalization ability of the method. Different from [14], we improve the generalization ability of the network by eliminating the distribution difference across the various operating conditions, instead of utilizing tricks and strategies during the training progress. To this end, we introduce a regularization term called center loss into the network objective function, which intentionally restricts the clustering property of the learned features. With the help of center loss, the network could minimize the domain difference across various seen operating conditions and improve the generalization ability on unseen operating conditions.

The main contributions of the proposed method can be summarized as follows:

1. We propose CNN-C network with the center loss to enhance the property of clustering within-class and the property of distinction among the classes.

2. For the testing samples under the seen operating conditions, CNN-C could not only encourage the features of the same class to cluster together but also improve the classification accuracy of different health conditions.

3. Without any information about the samples under the unseen operating conditions, the results show that the proposed method has an excellent generalization ability.

The organization of this paper is listed as follows. The detailed description of the CNN-C is introduced in Section II. The fault diagnosis framework is provided in Section III. In Section IV, the performance of the proposed method is evaluated in a bearing dataset. Finally, the conclusion and future work are discussed in Section V.

II. CNN WITH THE CENTER LOSS

The main advantage of the deep learning algorithm is automatically learning the nonlinear relationship between the original signals and class labels. Generally, a series of convolutional layers and max-pooling layers are combined to learn high-level representation, and then a classification layer is added to classify these abstract features. Based on the classification loss between the predicted label and the true label, the CNN network updates the weights in the convolutional layers by the back-propagation algorithm. Once the loss function is converged to a global minimum value, the network could identify the labeled data with high classification accuracy. However, this excellent classification performance is based on the assumption that the training samples and the testing samples are following the same distribution. The classification ability will decrease when faced with the different distribution between training and testing samples. Here is the situation: how to enhance the generalization ability of CNN without any information about the unseen testing samples.

A. THE CENTER LOSS

Recall the upper bounds of the generalization error in the statistical learning theory: Given a training set of \( \{(x_i, y_i)\}_{i=1}^{n} \) whose samples are drawn independent and identically distributed according to an unknown distribution \( D_n \), any function \( h \in H \) in the hypothesis space, with probability \( 1 - \delta \) at
least we have:

\[
R_T (h) \leq \hat{R}_S (h) + \frac{4}{n} \left[ \frac{\text{VC} (H) \log \frac{2en}{\text{VC} (H)} + \log \frac{4}{\delta}}{} \right] + d_H (D_S, D_t) + \lambda
\]  

(1)

where \( R_S (h) = \mathbb{E}_{x \sim D_t} [l (x, y)] \) is the expected risk of function \( h \) on the training set; \( \hat{R}_S (h) \) is the empirical risk; \( R_T (h) = \mathbb{E}_{x \sim D_t} [l (x, y)] \) is the expected risk of function \( h \) on the testing set; \( d_H (D_S, D_t) \) is the distance between the distribution of \( D_S \) and \( D_t; \lambda \) is a small value.

Hence, to increase the generalization error on the testing set, we should minimize the distance between the distributions of training set and testing set. In the fault diagnosis task on the unseen operating conditions, since there is no information about the shaft speed, normal samples or fault samples, it is unrealistic to estimate the distribution of the testing set. One of the possible solutions is to make full use of the seen operating conditions in the training set and learn the domain-invariant features.

In this paper, we introduce the center loss to the traditional convolutional neural network. By minimizing the intra-class variations, center loss cluster the learned features across various seen operating conditions. With the joint supervision of the center loss and the softmax loss, the network could seek a feature space across various seen operating conditions, and the learned features of the same class could minimize the domain difference while the features of different classes are discriminative and separable. In this way, the network shows a good generalization performance on the unseen operating condition.

The original intention of the center loss is to enhance the discriminative power in the deep learning method for face recognition [25]. The main idea of the center loss is simultaneously learning a center for features of each class and penalizing the distance between the features and their corresponding class centers. Therefore, the learned features have two essential properties: within-class clustered and between-class scattered. To minimize the within-class feature distance, the center loss function is intuitively proposed based on the sum distance between the features and the center point. For the given training dataset \( \{(x_i, y_i)\}_{i=1}^m \), the center loss can be computed as:

\[
\mathcal{L}_C = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|_2^2
\]  

(2)

where \( c_{y_i} \) is the center point of the learned features corresponding to the class label \( y_i \).

In this way, the center loss function controls the clustering property of the learned features. However, the center loss is not as smaller as better. Consider the following scenario: without the discriminative property, the center loss will force the network learning the same features regardless of the class labels. Therefore, it is necessary to combine the center loss with the classification loss to train CNN.

In theory, the entire training samples should be taken into consideration and the center point of each class should be computed by all the learned features in every iteration. To avoid inefficient progress, the center point is updated by the mini-batch, instead of the entire training samples [25]. The gradients of \( \mathcal{L}_C \) are computed as follows:

\[
\frac{\partial \mathcal{L}_C}{\partial x_i} = x_i - c_{y_i}
\]  

(3)

The update equation of \( c_{y_i} \) can be expressed as follows:

\[
\Delta c_j = \frac{\sum_{i=1}^m \delta (y_i = j) \cdot (c_j - x_i)}{1 + \sum_{i=1}^m \delta (y_i = j)}
\]  

(4)

where \( \delta (\text{condition}) = 1 \) if the condition is satisfied, and \( \delta (\text{condition}) = 0 \) if not.

Besides, a scalar parameter \( \alpha \) is used to control the learning rate of the centers [25]. In this paper, we set the scalar parameter as the default value \( \alpha = 0.5 \). In each iteration, the centers \( c_j \) for each \( j \) is updated as follows:

\[
c_j^{t+1} = c_j^t - \alpha \cdot \Delta c_j
\]  

(5)

B. NETWORK ARCHITECTURE

Based on the center loss of the learned features, the CNN-C architecture is proposed in this paper. The network structure contains a series of convolutional stages and a classification stage, as shown in Fig. 1.

In the convolutional stage, there is a convolutional layer, a batch normalization layer, a ReLU activation layer, and a max-pooling layer. The input samples of the network are normalized by zero mean value and one standard variance value. Besides, for the filters in the first convolutional stage, we set a wide kernel size. For filters in the other convolutional stages, the kernel size is chosen as a small value of \( 3 \times 1 \). The max-pooling size is set as \( 2 \times 1 \). For the final layer in the last convolutional stage, the nodes are flattened and fed into the softmax output layer.

In general, the fault diagnosis task of the rotating machinery is modeled as a multiclass classification problem. The features are learned by the former convolutional stages. To ensure the features clustered within-class and scattered between-class, the center loss is added in the base of classification cross-entropy loss. Then the network parameters of each layer in the CNN-C are trained by minimizing the loss function with Adam optimization algorithm. Formally, the final objective function of CNN-C is defined as follows:

\[
\mathcal{L}_{\text{CNN-C}} = \mathcal{L}_S + \lambda \mathcal{L}_C
\]  

(6)

where \( \mathcal{L}_S \) represents the softmax loss. \( \mathcal{L}_C \) represents the center loss defined in Eq. (1). \( \lambda \) is a trade-off parameter that adjusts the proportion of the center loss in the classification loss.

The softmax loss can be defined as follows.

\[
\mathcal{L}_S = - \sum_{i=1}^m \log \frac{e^{W_{yi} x_i + b_y}}{\sum_{j=1}^n e^{W_{ji} x_i + b_j}}
\]  

(7)
FIGURE 1. The overall architecture of the proposed method.

where \( x_i \) is the \( i \)-th feature belonging to the \( y_i \) class. \( W_j \) and \( b_j \) is the \( j \)-th column of the weights and bias in the last fully connected layer, respectively. \( m \) is the size of mini-batch and \( n \) is the number of classes.

III. THE PROPOSED FAULT DIAGNOSIS METHOD

A. FAULT DIAGNOSIS FRAMEWORK

Based on the proposed CNN with the center loss, a novel fault diagnosis framework is designed for the rotating machinery to achieve acceptable diagnosis performance both on seen and unseen operating conditions. The flowchart of the framework is shown in Fig 2. The main steps of the fault diagnosis method can be summarized as follows:

Step 1: Data Acquisition: Based on the existing operating conditions, sensors are installed to collect the vibration signals under different health conditions. The signals are sliced into a series of samples and these samples consist of the training database. For each sample, it should be normalized by zero mean and unit variance.

Step 2: Model Train: The samples in the training database are employed to train the fault diagnosis model. First, the Bayesian Optimization method with cross-validation is applied to select the appropriate architecture parameters, including the filter number, filter size. Then, the center loss parameter \( \lambda \) is determined with the trade-off computing time and training classification accuracy.

Step 3: Fault Diagnosis: To evaluate the performance of the proposed method, the testing samples are collected from the seen operating conditions and unseen operating conditions, respectively. In both of the fault diagnosis scenarios, the testing samples should be normalized by zero mean and unit variance before feeding into the trained CNN-C for condition identification.

B. NETWORK PARAMETERS

The network adopted in this work has three convolutional and pooling stages. Five network hyper-parameters including the number of filters in each convolutional layer, filter size and stride size of the first convolutional layer, are selected using the Bayesian Optimization methods by 3-fold cross-validation on the training data set. For each convolutional layer, the filter number ranges from 16 to 64. The filter size of the first convolutional layer ranges from 8 to 512, and the corresponding stride size is set as \([1/2, 1/4, 1/8]\) of the filter size. We set 100 possible hyper-parameters combination in the Bayesian Optimization process. The optimal network can be obtained as listed in Table 1. Besides, the learning rate is 0.005.

| Layer Name  | Filter number | Filter size | Stride size | Activation |
|-------------|---------------|-------------|-------------|------------|
| Convolution | 64            | 64×1        | 8×1         | ReLU       |
| Max pooling | -             | 2×1         | 2×1         |            |
| Convolution | 64            | 3×1         | 1×1         | ReLU       |
| Max pooling | -             | 2×1         | 2×1         | -          |
| Convolution | 64            | 3×1         | 1×1         | ReLU       |
| Max pooling | -             | 2×1         | 2×1         | -          |
| Flatten     | -             | -           | -           | -          |
| Classification | 7            | -           | -           | Softmax    |

It should be noted that the center loss parameter \( \lambda \) is determined based on the optimal network, which is a trade-off between the classification loss and the center loss. When \( \lambda \) is small, the whole loss of the network is dominated by the classification loss, after convergence, the center loss would remain a relatively large value. The loss function is mainly decreased by softmax loss. Once the loss function is converged, the learned feature is discriminative but scattered. In contrast, with a large value, the loss function is decreased with the center loss. The loss function forces the learned features clustering together but ignores the difference among different health conditions, which might lead to a poor classification result. Therefore, it is necessary to make a trade-off between the classification loss and center loss, which makes the network learn clustered and discriminative features. The detailed discussion of the center loss parameter is demonstrated in Section VI.
C. PROPERTIES OF THE PROPOSED METHOD

In this paper, we mainly concentrate on promoting the generalization performance of CNN for rotating machinery fault diagnosis by learning class-clustered features. Fault diagnosis is always tough when the training samples and testing samples follow a different distribution, and the hardest part is that there is no information about the unseen operating conditions (such as the shaft speed, normal condition samples). To tackle this problem, one of the feasible solutions is encouraging the network to learn the features with less dependency on the operating conditions. As a result, the various operating conditions have few effects on the feature distribution, and the learned features have sufficient generalization ability to deal with the unseen operating conditions. To this end, we introduce the center loss combined with the classification loss function. The learned features are discriminative enough to classify different health conditions and clustered enough to eliminate the effects of the various operating conditions. In summary, the main properties of our proposed CNN-C fault diagnosis framework can be described as follows:

1. Discriminative: The learned features are discriminative between-class. Based on the softmax loss function, the network could converge at a small classification loss and achieve a good classification performance. In the fault diagnosis field, the method is good at dealing with different fault types and diverse fault severity levels.

2. Clustered: The learned features are clustered within-class. With the help of the center loss, the network is encouraged to cluster the features of the same class. In the fault diagnosis scenario, the learned features of the same health condition under different operating conditions are clustered together. Therefore, the operating conditions have little influence on the feature distribution and the network is able to learn the common fault characteristics among these operating conditions.

3. General: The proposed method is able to deal with the fault diagnosis tasks under unseen operating conditions. Most deep learning research works focus on the scenario that the training samples and testing samples follow the same distribution, or the shaft speed and normal samples under the unseen operating conditions are needed to fine-tuning the pre-trained network. On the contrary, this work is mainly to solve the fault diagnosis tasks without any information about the unseen operating conditions. The discriminative and
clustered features could enhance the generalization ability of the network.

IV. EXPERIMENTAL VERIFICATION

The proposed method is evaluated on two diagnosis cases, roller bearing and gearbox. The tasks of diagnosing the samples from both seen operating conditions (scenario A) and unseen operating conditions (scenario B) with respect to training data are organized. In specific, scenario A denotes the situation that the test samples are collected from the operating conditions included by the training data. Scenario B denotes the opposite situation. Traditional CNN and shallow methods are also compared to demonstrate the superior performance of the proposed method.

A. EXPERIMENTAL SETUP

1) DATA DESCRIPTION

There are two datasets to verify the proposed method, including a bearing dataset and a gearbox dataset.

In the bearing dataset, the experiment data is collected on the bearing test rig shown in Fig. 3. The experimental rig mainly includes a driving motor, a conveyor belt, a loading system, a shaft, and two bearings. The acceleration sensor is installed on the bearing pedestal to collect the vibration signals in the vertical direction. The sampling frequency is set as 20 KHz. The vibration signals are collected under seven health conditions: normal; roller with pitting fault; roller with groove fault; roller with wear failure; roller with missing piece fault; inner race with wear fault; outer race with breakage fault, as shown in Fig. 4. There are 15 operating conditions in the bearing test, with the speeding ranged from 100 r/min to 1500 r/min. For each health condition under the single operating condition, the vibration signals are sliced into a series of samples with the length of 2048. The total samples are divided into training samples and testing samples. The detailed description of the samples is listed in Table 2. In scenario A1, both the training and testing samples contain 15 operating conditions, following the same distribution. In scenario B1, the training samples contain 14 operating conditions, and the testing samples consist of the rest operating condition as the unseen operating condition.

In the gearbox dataset, the spur gearbox data is from the 2009 challenge data of Prognostics and Health Management (PHM) society [26]. The vibration signals are collected by two accelerometers installed on the input shaft end and output shaft end. The sampling frequency is 66.67 kHz, and the shaft speed ranges from 30 Hz to 50 Hz, corresponding to the high load and low load, respectively. There are eight health conditions with compound faults testing in the experiment. Similar to the bearing dataset, the samples are divided into the training samples and testing samples with the length of $N = 2048$, as listed in Table 3. Same as the bearing dataset,
the training and testing samples are both composed of 10 operating conditions in scenario A2. In contrast, training samples in scenario B2 includes 9 operating conditions, while the rest operating condition is chosen as the unseen operating condition for the testing samples.

2) COMPARISON METHODS
To show the superior performance of the proposed method, several methods are selected as comparison methods, including traditional CNN and shallow methods. The architecture of CNN is the same as the proposed methods, and the results are averaged by 10 trials to reduce the randomness. Meanwhile, the handcrafted features are extracted both in the time domain and in the frequency domain. There are 12 features in the time domain, including peak, peak-to-peak, absolute mean, square root, standard variance, root mean square, waveform index, margin index, peak index, impulse factor, skewness index, and kurtosis index. Moreover, there are 6 features in the frequency domain, consisting of frequency center, root mean square frequency, standard deviation frequency, and CP1, CP2, CP3 [27], [28]. These statistical features are fed into the shallow methods to classify the health conditions. Random forest (RF), Support vector machine (SVM) and K-Nearest Neighbor (KNN) are utilized in this paper since the excellent performance in kinds of fault diagnosis applications.

B. EFFECTS OF CENTER LOSS
The center loss parameter $\lambda$ controls the proportion of the center loss in the total loss function. It should be noted that the center loss is encouraged to employ with the softmax loss simultaneously. Without center loss, the features with the same class are scattered. In the fault diagnosis field, the features learned by the network are discriminative enough to distinguish different health conditions, but the features of the same health condition are various under different operating conditions. If the operating conditions are changed, the learned features and discriminative manner could not be generalized to other different but unseen operating conditions. On the other hand, classification loss is also requisite. Without it, the features learned from different operating conditions and classes would cluster together. Then the network is meaningless without the classification ability. Therefore, a proper value of $\lambda$ could not only guarantee the classification ability of different health conditions but also make sure the features are cluster under different operating conditions. Since the learned features without the dependency on the operating conditions, it could show a good performance on the fault diagnosis tasks under various operating conditions. In other words, the network has a generalization ability to deal with unseen operating conditions.

To investigate the effect of the center loss parameter, different values of $\lambda$ are implied in the bearing dataset, scenario A1. The classification accuracies with different center loss parameters after 200 epochs are listed in Table 4. The center loss, softmax loss, and accuracy in the training progress are shown in Fig.5. The results can be summarized as follows: First, for a small value of center loss parameter ($\lambda = 0, 1e-6, 1e-5$), the softmax loss is rapidly decreasing with the increase epochs. The classification accuracy converges to around 96%. However, the center loss remains a large value, which indicates that the features of the same class are far away from the cluster center point. Second, with the rise of the parameter ($\lambda = 1e-4, 1e-3$), the softmax loss converges slowly and the center loss is steady at a small value. This shows that the features are still discriminative among the different health conditions, and tend to be clustered in the same health condition. In other words, the network keeps the high classification ability and increases the generalization ability, based on the sacrifice of the training time. Third, if the parameter is too large ($\lambda = 1e-2, 1e-1$), the loss function focus on reducing the center loss and regard the softmax loss. The softmax loss could not converge to the minimum value and the classification accuracy rapidly drops. Especially, the network has no classification ability with an accuracy of 14.24%. In this situation, the learned features of different health conditions and operating conditions are clustered together, which are useless and meaningless.

| Class Label | Shaft speed (Hz) | Load | Training samples | Testing samples |
|-------------|-----------------|------|-----------------|----------------|
| 1           | 30/35/40/45/50  | Low/High | 1300  | 1300  |
| 2           | 30/35/40/45/50  | Low/High | 1300  | 1300  |
| 3           | 30/35/40/45/50  | Low/High | 1300  | 1300  |
| 4           | 30/35/40/45/50  | Low/High | 1300  | 1300  |
| 5           | 30/35/40/45/50  | Low/High | 1300  | 1300  |
| 6           | 30/35/40/45/50  | Low/High | 1300  | 1300  |
| 7           | 30/35/40/45/50  | Low/High | 1300  | 1300  |
| 8           | 30/35/40/45/50  | Low/High | 1300  | 1300  |

Besides, to have a better understanding of the center loss effects on the learned features, t-SNE [29] is carried out to reduce the feature dimension and visualize the feature distribution. The results are shown in Fig.5, different colors correspond to different health conditions and different numbers correspond to different operating conditions. It should be noted that the network is the traditional CNN when the parameter is equal to zero. From the figure, we could see that: First, although traditional CNN has a high classification accuracy, the features within the same class are separated.
from each other. Different operating conditions correspond to different feature distribution areas, which means that a poor generalization ability. Therefore, traditional CNN is only available to deal with the testing samples under the seen operating condition. Second, with the increase of the center loss parameter, the features of different health conditions are discriminative while the features of the same health condition under different operating conditions are clustered together. This shows that the features are effective to classify the health conditions and insensitive to the operating conditions. The network has good generalization ability. Third, with a large value of $\lambda = 0.01$, the features of different health conditions overlap with each other. One of the possible reasons is that the large center loss parameter force the loss function decreasing the center loss and ignoring the softmax loss. The network trends to converge into a cluster of features, regardless of the health conditions and operating conditions.

Based on the analysis above, the selection of the center loss parameter is making the features of the same health condition under various operating conditions as clustered as possible and ensuring the difference among different health conditions. Moreover, the training time is another key point to consider. Therefore, we set $\lambda = 1e-4$ in the bearing dataset. In the same way, the center loss parameter $\lambda$ in the gearbox dataset is chosen as $\lambda = 1e-3$.

### C. PERFORMANCE ON THE SEEN OPERATING CONDITIONS

In scenario A1 and A2, the training samples and testing samples are collected under the same operating conditions, which evaluates the classification performance of the fault diagnosis framework on the seen operating conditions.

To compare with the results of the proposed method, RF classifier, KNN classifier, and SVM classifier are employed to the datasets with 18 handcrafted statistical features. The hyper-parameters of these shallow methods are selected by a grid search algorithm, as listed in Table 5. Meanwhile, the traditional CNN with the same architecture is also tested. The classification accuracies of all the methods are listed in Table 6. Note that the deep learning methods are averaged by 10 trails to eliminate the randomness.

From the classification results in scenario A1, we could observe that: First, the traditional handcrafted features based on the expert knowledge and signal processing skills show poor performance on the bearing dataset. Although the RF classifier achieves 100% training accuracy, the testing classification accuracy is around 80.63%. Therefore, it is difficult to diagnosis the bearing fault under various operating conditions based on the handcrafted features. Second, traditional CNN benefits from the powerful feature learning ability and achieves 96.30% testing accuracy. Compared with the shallow methods, the deep learning method is able to map the nonlinear relationship between the original signals and class labels. The network could automatically learn the remarkable features and classify different health conditions. Third, with the advantages of the center loss, the proposed CNN-C method forces the features of the same class clustering together. The features are not scattered under different operating conditions. The property of the clustering makes the network classify different health conditions.
more easily. Hence, CNN-C has the highest testing classification of 97.20% among the methods.

The confusion matrices of different methods are shown in Fig. 7 and Fig. 8. In scenario A1, it is obvious that the shallow method has poor classification ability to distinguish class label 2-5. One of the possible reasons is that these classes are all the faults that occurred on the roller of the bearing. These faults are characterized by similar demodulated components and low-frequency impulse components. The handcrafted features based on expert knowledge could not extract the fault characteristics among these roller faults. In contrast, deep learning methods automatically learn the features and outperform shallow methods. The network not only reduces human labor but also enhances the classification performance. Moreover, CNN-C introduces the center loss to the loss function. The features of the same class cluster together and the classification accuracy of each class has an obvious improvement.

D. PERFORMANCE ON THE UNSEEN OPERATING CONDITIONS

To evaluate the generalization ability of the proposed method, the operating conditions of the testing samples are different and unseen compared with the training samples. There are 15 and 10 types of unseen operating conditions in scenario B1 and B2, respectively. The classification accuracies of different methods are listed in Table 7 and Table 8. The highest classification accuracy among these methods is marked as bold.

Based on table 7, the scenario B1 of the bearing dataset, the following results are summarized. First, the shallow methods show a poor generalization performance in the bearing...
Y. Yang et al.: Learn Generalization Feature via Convolutional Neural Network

**FIGURE 8.** The confusion matrices of different methods in scenario A2: (a) SVM; (b) CNN; (c) CNN-C.

**TABLE 7.** The classification accuracies of different methods in scenario B1 (bearing dataset under the unseen operating conditions).

| Testing operating condition | KNN       | RF        | SVM       | CNN       | CNN-C     |
|-----------------------------|-----------|-----------|-----------|-----------|-----------|
| 100                         | 22.86     | 24.06     | 28.17     | **40.27±1.80** | 37.92±1.32 |
| 200                         | 31.92     | 27.90     | 41.83     | **75.23±1.88** | 73.13±1.40 |
| 300                         | 40.54     | 39.51     | 49.60     | 90.79±0.67 | **92.02±0.70** |
| 400                         | 44.15     | 45.00     | 53.53     | 95.39±0.99 | **96.62±0.48** |
| 500                         | 43.71     | 50.27     | 54.20     | 95.46±0.77 | **97.07±0.68** |
| 600                         | 48.97     | 55.67     | 63.97     | 96.34±0.58 | **96.76±0.96** |
| 700                         | 63.04     | 66.92     | 75.98     | 98.40±0.39 | **98.44±0.39** |
| 800                         | 70.76     | 74.46     | 82.77     | 98.60±0.25 | **98.92±0.39** |
| 900                         | 74.60     | 80.27     | 88.57     | **99.32±0.15** | 99.17±0.45 |
| 1000                        | 82.86     | 86.29     | 91.88     | 99.55±0.17 | **99.79±0.13** |
| 1100                        | 85.58     | 88.88     | 95.09     | 99.71±0.15 | **99.80±0.15** |
| 1200                        | 89.38     | 92.05     | 96.83     | **99.65±0.24** | 99.56±0.18 |
| 1300                        | 91.47     | 92.81     | 97.54     | 98.91±0.68 | **99.62±0.20** |
| 1400                        | 92.05     | 93.17     | 96.88     | 99.11±0.34 | **99.34±0.19** |
| 1500                        | 86.88     | 83.39     | 94.20     | 97.71±0.38 | **98.61±0.29** |

**TABLE 8.** The classification accuracies of different methods in scenario B2 (gearbox dataset under the unseen operating conditions).

| Testing operating condition | KNN       | RF        | SVM       | CNN       | CNN-C     |
|-----------------------------|-----------|-----------|-----------|-----------|-----------|
| 30H                         | 48.17     | 49.33     | 55.48     | 81.86±1.90 | **83.15±1.83** |
| 30L                         | 39.90     | 44.71     | 53.17     | 81.08±3.24 | **83.38±3.01** |
| 35H                         | 66.15     | 77.98     | 78.27     | 94.24±1.34 | **95.95±1.16** |
| 35L                         | 55.19     | 58.94     | 68.56     | **95.72±1.56** | 95.69±2.56 |
| 40H                         | 72.79     | 74.42     | 76.15     | 94.19±1.46 | **95.88±1.32** |
| 40L                         | 51.73     | 68.08     | 59.90     | 86.74±5.46 | **88.95±3.32** |
| 45H                         | 73.17     | 71.44     | 76.35     | 91.76±1.66 | **94.30±0.93** |
| 45L                         | 64.62     | 67.02     | 59.42     | 76.92±3.97 | **82.88±4.07** |
| 50H                         | 55.19     | 47.40     | 46.15     | 83.95±3.30 | **86.23±2.74** |
| 50L                         | 56.92     | 51.25     | 48.56     | 79.75±3.96 | **81.32±4.61** |

dataset. The handcrafted features of different health conditions are overlapped together, and the features of the same health condition are far away from each other under different operating conditions. Therefore, the handcrafted features are not suitable to deal with the unseen operating conditions. **Second**, CNN has a better generalization performance compared with the shallow methods. The learned features are helpful to classify testing samples when the seen operating conditions are similar to the unseen operating condition. For instance, the feature distribution of the unseen 1000r/min is similar to the seen 900r/min and 1100r/min. If we were able to diagnose the samples under 900r/min and 1100r/min precisely, the fault diagnosis framework would perform well dealing with the unseen 1000r/min samples. **Third**, the CNN-C has the best generalization ability to deal with the unseen operating condition. With the help of the center loss, CNN-C could not only learn the discriminative features of different health conditions but also encourage these features of the same class cluster together, as shown in Fig.9. The center loss estimates the effects of various operating conditions on the feature distribution.

**FIGURE 9.** The feature distribution in scenario B1 (bearing dataset under the unseen operating conditions of 1300 r/min).
Besides, the classification accuracies of lower speed conditions in the bearing dataset are always smaller than the higher speed conditions. The classification accuracies are 37.92% and 75.23% under the testing operating conditions of 100r/min and 200r/min, respectively, while the classification accuracies achieve to about 99% under the higher speed conditions. One of the possible reasons is that the sampling frequency is too high ($f_s = 20$ kHz) and the data length of the sample is too short ($N = 2048$). If we transform the signals in the time domain to the angle domain, the roller bearing only rotate 0.1707 circles when the operating condition of 100r/min. With the consideration of the fault frequency in the roller bearing, the fault characteristics are not obvious enough in the limited data length. It is difficult to extract useful features under the slow speed operating conditions. With the speed increase of the testing operating conditions, the bearing rotates more angles and the fault characteristics are clearer. Especially, the samples contain a complete rotating circle of the bearing with a speed of 600r/min. The testing classification accuracies achieve to around 97%, which means that the network is able to extract the remarkable and effective features to deal with the unseen operating conditions.

As for Table 8, the scenario B2 of the gearbox dataset, the results are similar to the bearing dataset. The shallow methods based on the handcrafted features are not suitable to deal with the fault diagnosis tasks under the unseen operating conditions. The traditional CNN has an improvement on the unseen operating conditions. This means that the learned features based on deep learning algorithms have a better generalization ability than the handcrafted features. As shown in Fig.10, the feature distribution of testing samples under unseen operating conditions (marked as solid triangles) overlap with the training samples (marked as hollow circles). The center loss encourages the features of the same class cluster together and the classification loss keeps the individual variation among different health conditions. In this way, the loss function enhances the generalization ability of the network and remains the classification ability. The proposed method CNN-C outperforms the traditional CNN and shallow methods.

It should be noted that the classification accuracies of the low (83.15% and 83.38% for 30 Hz shaft speed) and high (86.23% and 81.32% for 50 Hz shaft speed) speeds are smaller than the medium speeds (95.95%, 95.88, and 94.30% for 35-45 Hz shaft speed). One of the possible reasons is that the feature distribution under the unseen operating conditions is similar to the nearest seen operating conditions. Since the training samples contains the shaft speeds below and above the medium speeds of unseen operating conditions, the network could learn more universal features and show a better generalization ability on these unseen operating condition.

Besides, the computational time of each epoch in the training progress is also considered. The computational time is averaged by 150 trials to avoid the randomness. The epoch time of CNN and CNN-C is 3.24 s and 3.44 s, respectively.

The results show that center loss has no obvious effect on computational efficiency.

Based on the comparison results of shallow methods and traditional CNN in the bearing dataset and gearbox dataset, the proposed method is able to enhance the generalization ability of the network and achieves the best performance. Therefore, it is promising to deal with fault diagnosis tasks under the unseen operating conditions.

V. CONCLUSION

In this paper, we mainly discuss the generalization ability of CNN to deal with the unseen operating conditions in the fault diagnosis field. We introduce the center loss to the traditional CNNs and build an end-to-end fault diagnosis framework. Then, we design two scenarios in a bearing dataset and a gearbox dataset, based on the testing samples under the seen and unseen operating conditions, respectively. The tasks are employed to evaluate the classification ability and generalization ability of the proposed method.

The experimental results show that the proposed CNN-C method could not only classify different health conditions on the seen operating conditions but also perform well on the unseen operating conditions. A proper center loss parameter enhances the generalization ability of the network and remains the classification ability. It is a promising tool to deal with the fault diagnosis task of the rotating machinery since most operating conditions in the real world are unseen.

In future work, more real-world datasets and start-of-art convolution neural networks will be applied to verify the effectiveness of the classification ability and generalization ability. Moreover, advanced techniques such as weight orthogonal will be introduced to the network.
We appreciate the Prognostics and Health Management (PHM) society for the 2009 challenge dataset.

REFERENCES

[1] J. Yin, M. Xu, and H. Zheng, “Fault diagnosis of bearing based on symbolic aggregate approXimation and Lempel-Ziv,” *Measurement*, vol. 138, pp. 206–216, May 2019.

[2] Y. Yang, H. Zheng, Y. Chen, J. Yin, M. Xu, and Y. Chen, “Refined composite multivariate multiscale symbolic dynamic entropy and its application to fault diagnosis of rotating machine,” *Measurement*, vol. 151, Feb. 2020, Art. no. 107233.

[3] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, “An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data,” *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.

[4] G. Jiang, H. He, J. Yan, and P. Xie, “Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox,” *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3196–3207, Apr. 2019.

[5] 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.

[6] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, “Real-time motor fault detection by 1-D convolutional neural networks,” *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016.

[7] O. Janssens, V. Slavkovicj, B. Vervisch, K. Stockman, M. Loccuferi, S. Verstockt, R. Van de Walle, and S. Van Hoecke, “Convolutional neural network based fault detection for rotating machinery,” *J. Sound Vibrat.*, vol. 377, pp. 331–345, Sep. 2016.

[8] Y. Yang, H. Zheng, Y. Li, M. Xu, and Y. Chen, “A fault diagnosis scheme for rotating machinery using hierarchical symbolic analysis and convolutional neural network,” *ISA Trans.*, vol. 91, pp. 235–252, Aug. 2019.

[9] F. Jia, Y. Lei, N. Lu, and S. Xing, “Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization,” *Mech. Syst. Signal Process.*, vol. 110, pp. 349–367, Sep. 2018.

[10] S. Haidong, C. Junsheng, J. Hongkai, Y. Yu, and W. Zhanhao, “Enhanced deep gated recurrent unit and complex wavelet packet energy moment entropy for early fault prognosis of bearing,” *Knowl.-Based Syst.*, vol. 188, Jan. 2020, Art. no. 105022.

[11] Z. He, H. Shao, X. Zhang, J. Cheng, and Y. Yang, “Improved deep transfer auto-encoder for fault diagnosis of gearbox under variable working conditions with small training samples,” *IEEE Access*, vol. 7, pp. 115368–115377, 2019.

[12] J. Wang, R. Zhao, and R. X. Gao, “Probabilistic transfer factor analysis for machinery autonomous diagnosis cross various operating conditions,” *IEEE Trans. Instrum. Meas.*, Jan. 3, 2020, doi: 10.1109/TIM.2019.2963731.

[13] H. Zheng, R. Wang, Y. Yang, J. Yin, Y. Li, Y. Li, and M. Xu, “Cross-domain fault diagnosis using knowledge transfer strategy: A review,” *IEEE Access*, vol. 7, pp. 129260–129290, 2019.

[14] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, “A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load,” *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, Feb. 2018.

[15] T. Han, C. Liu, L. Wu, S. Sarkar, and D. Jiang, “An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems,” *Mech. Syst. Signal Process.*, vol. 117, pp. 170–187, Feb. 2019.

[16] T. Han, C. Liu, W. Yang, and D. Jiang, “Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application,” *ISA Trans.*, vol. 97, pp. 269–281, Feb. 2020.

[17] T. Han, C. Liu, W. Yang, and D. Jiang, “Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions,” *ISA Trans.*, vol. 93, pp. 341–353, Oct. 2019.

[18] T. Han, C. Liu, W. Yang, and D. Jiang, “A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults,” *Knowl.-Based Syst.*, vol. 165, pp. 474–487, Feb. 2019.

[19] Y. Chen, G. Peng, C. Xie, W. Zhang, C. Li, and S. Liu, “ACDIN: Bridging the gap between artificial and real bearing damages for bearing fault diagnosis,” *Neurocomputing*, vol. 294, pp. 61–71, Jun. 2018.

[20] Z. Zhu, G. Peng, Y. Chen, and H. Gao, “A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis,” *Neurocomputing*, vol. 323, pp. 62–75, Jan. 2019.

[21] P. Peng and J. Wang, “NOSCNN: A robust method for fault diagnosis of RV reducer,” *Measurement*, vol. 138, pp. 652–658, May 2019.

[22] L. Wen, L. Gao, and X. Li, “A new snapshot ensemble convolutional neural network for fault diagnosis,” *IEEE Access*, vol. 7, pp. 32037–32047, 2019.

[23] J. Wang, Y. Ma, Z. Huang, R. Xue, and R. Zhao, “Performance analysis and enhancement of deep convolutional neural network,” *Bus. Inf. Syst. Eng.*, vol. 61, pp. 311–326, Mar. 2019.

[24] D. Wei, K. Wang, S. Heyns, and M. J. Zuo, “Convolutional neural networks for fault diagnosis using rotating speed normalized vibration,” in *Proc. Int. Conf. Condition Monit. Machinery Non-Stationary Oper.*, Santander, Spain, Jun. 2018, pp. 67–76.

[25] S. Soetens, A. Sarris, and K. Vansteenbussche, “A discriminative feature learning approach for deep face recognition,” in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 499–515.

[26] PHM Society. *PHM09 Gearbox Dataset*. Accessed: May 2019. [Online]. Available: https://www.phmsociety.org/competition/PHM09

[27] H. Zheng, R. Wang, Y. Yang, Y. Li, and M. Xu, “Intelligent fault identification based on multisource domain generalization towards actual diagnosis scenario,” *IEEE Trans. Ind. Electron.*, vol. 67, no. 2, pp. 1293–1304, Feb. 2020.

[28] H. Zheng, R. Wang, J. Yin, Y. Li, H. Lu, and M. Xu, “A new intelligent fault identification method based on transfer locality preserving projection for actual diagnosis scenario of rotating machinery,” *Mech. Syst. Signal Process.*, vol. 135, Jan. 2020, Art. no. 106344.

[29] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

YUAN TAO YANG received the B.S. and M.S. degrees in mechanics from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Deep Space Exploration Research Center.

His research interests include fault diagnosis of machinery, intelligent fault diagnosis method, and deep learning.

JIAN CHENG YIN received the B.S. degrees in mechanics from the Harbin Institute of Technology, Harbin, China, in 2014, where he is currently pursuing the Ph.D. degree in mechanics with the Deep Space Exploration Research Center.

His research interests include fault diagnosis of machinery, residual life prediction method, and signal processing.

HUAILIANG ZHENG received the B.S. and M.S. degrees in mechanics from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Deep Space Exploration Research Center.

His research interests include fault diagnosis of machinery, intelligent fault diagnosis method, and transfer learning.
YUQING LI received the B.E. degree in mechanical design manufacturing and automation from the Harbin Institute of Technology, Harbin, China, in 2002, and the M.E. and Ph.D. degrees in general mechanics from the Harbin Institute of Technology, in 2004 and 2008, respectively.

He is currently an Associate Professor with the Harbin Institute of Technology. His main research interests are planning and scheduling of spacecraft, and spacecraft fault detection and diagnosis.

MINQIANG XU received the B.E. degree in electronics from Peking University, Beijing, China, in 1983, the M.E. degree in nuclear physics from Northeast Normal University, Changchun, China, in 1989, and the Ph.D. degree in general and fundamental mechanics from the Harbin Institute of Technology, Harbin, China, in 1999.

Since 2000, he has been a Professor with the Harbin Institute of Technology. His research interests include machinery and spacecraft fault detection and diagnosis, signal processing, and space debris modeling.

YUSHU CHEN received the B.E. degree in mechanics from Tianjin University, Tianjin, China, in 1956, and the Ph.D. degree from the Mechanical Institute, Soviet Academy of Sciences, in 1963.

He is currently a Professor with the Harbin Institute of Technology. He is particularly interested in non-linear dynamics.

* * *