


Article

A Novel Multi-Task Self-Supervised Transfer Learning Framework for Cross-Machine Rolling Bearing Fault Diagnosis

Lujia Zhao ¹, Yuling He ^{2,*} , Derui Dai ², Xiaolong Wang ², Honghua Bai ³ and Weiling Huang ³

¹ Engineering Training and Innovation and Entrepreneurship Education Center, North China Electric Power University, Baoding 071003, China; 40852095@ncepu.edu.cn

² Department of Mechanical Engineering, North China Electric Power University, Baoding 071003, China; 120232102035@ncepu.edu.cn (D.D.); wangxiaolong0312@ncepu.edu.cn (X.W.)

³ Zhejiang Zhenxing Axiang Group, Huzhou 313000, China; benjamin.bai@axiang.com (H.B.); huang.weiling@axiang.com (W.H.)

* Correspondence: heyuling1@ncepu.edu.cn

Abstract: In recent years, intelligent methods based on transfer learning have achieved significant research results in the field of rolling bearing fault diagnosis. However, most studies focus on the transfer diagnosis scenario under different working conditions of the same machine. The transfer fault diagnosis methods used for different machines have problems such as low recognition accuracy and unstable performance. Therefore, a novel multi-task self-supervised transfer learning framework (MTSTLF) is proposed for cross-machine rolling bearing fault diagnosis. The proposed method is trained using a multi-task learning paradigm, which includes three self-supervised learning tasks and one fault diagnosis task. First, three different scales of masking methods are designed to generate masked vibration data based on the periodicity and intrinsic information of the rolling bearing vibration signals. Through self-supervised learning, the attention to the intrinsic features of data in different health conditions is enhanced, thereby improving the model's feature expression capability. Secondly, a multi-perspective feature transfer method is proposed for completing cross-machine fault diagnosis tasks. By integrating two types of metrics, probability distribution and geometric similarity, the method focuses on transferable fault diagnosis knowledge from different perspectives, thereby enhancing the transfer learning ability and accomplishing cross-machine fault diagnosis of rolling bearings. Two experimental cases are carried out to evaluate the effectiveness of the proposed method. Results suggest that the proposed method is effective for cross-machine rolling bearing fault diagnosis.

Keywords: fault diagnosis; rolling bearing; self-supervised learning; transfer learning; multi-task learning



Citation: Zhao, L.; He, Y.; Dai, D.; Wang, X.; Bai, H.; Huang, W. A Novel Multi-Task Self-Supervised Transfer Learning Framework for Cross-Machine Rolling Bearing Fault Diagnosis. *Electronics* **2024**, *13*, 4622. <https://doi.org/10.3390/electronics13234622>

Academic Editor: Davide Astolfi

Received: 14 October 2024

Revised: 12 November 2024

Accepted: 15 November 2024

Published: 23 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Rolling bearings, as crucial components of rotating machinery, directly affect the stability of equipment operation. Often operating under variable loads and complex environments, rolling bearings are prone to faults. These faults can lead to economic losses at best and safety accidents at worst [1–3]. Therefore, fault diagnosis of rolling bearings is of great value for ensuring the safe operation of rotating machinery. With the continuous development of deep learning, intelligent fault diagnosis technology has achieved significant results in the field of rolling bearing fault diagnosis and has become a hot research topic. Most deep learning-based rolling bearing fault diagnosis methods generally assume that the probability distributions of the training and testing datasets are similar [4,5]. However, in practical situations, the operating conditions of rolling bearings are time-variant, and the distributions of data collected under time-variant conditions are different. Consequently, a model that performs well on data from one operating condition may perform poorly on data from another condition.

Transfer learning can acquire generalized and transferable fault diagnosis knowledge, enabling fault diagnosis of rolling bearings under different working conditions without training models from scratch using new data or fine-tuning the models with a small quantity of new data [6]. Domain adaptation, as a commonly used transfer learning method, is widely applied in the field of fault diagnosis [7]. Wasserstein distance, as a commonly used probability distribution measurement method, has been widely studied in the field of domain adaptive rolling bearing fault diagnosis [8]. Wang et al. [9] proposed a new cross domain fault diagnosis method based on conditional Wasserstein distance, which improved the model's ability to transfer fault diagnosis. To solve the problem of bearing transfer fault diagnosis from multiple domains to the target domain, Zhao et al. [10] proposed a transfer learning approach called conditional weighting transfer Wasserstein auto encoder. By utilizing Wasserstein distance to generate transferable features that approximate Gaussian distributions, the ability of the model to diagnose transfer faults has been improved. However, Wasserstein distance only narrows down the difference between the source and target domains from the perspective of probability distribution, and a single measurement method may not fully exploit transferable features when facing more complex cross-mechanical fault diagnosis problems. Cosine similarity, as a measurement method for mining similarity relationships between data, has been applied in fields such as image recognition, neural network optimization, and computer recommendation systems [11–13]. Therefore, combining Wasserstein distance and cosine similarity, a multi-perspective feature transfer method is proposed to solve domain adaptation.

Although the aforementioned domain adaptation methods can achieve fault diagnosis of bearings under different working conditions, the data used for model training all come from the same machine. In practical applications, data usually come from multiple machines of the same type or different types of machines. The differences between data also need to consider the impact of objective factors such as machine characteristics and operating environments [14,15]. Compared with data obtained from different working conditions of the same machine, the distribution differences in data collected from different machines are greater, and the negative impact on the model is greater. Some studies have attempted to solve the problem of cross-machine rolling bearing fault diagnosis. Yang et al. [16], inspired by the idea of transfer learning, proposed a domain adaptation transfer neural network for transferring diagnostic knowledge from laboratory datasets to actual locomotive bearings, achieving bearing transfer fault diagnosis. Guo et al. [17] proposed a deep convolutional transfer learning network based on transfer learning for transfer fault diagnosis between different machines, improving the model's transfer learning ability by maximizing domain recognition errors and minimizing probability distribution distances. To improve the domain generalization ability of the model, Jia et al. [18] used defined causal and noncausal factors to train a fault diagnosis model. Jia et al. [19] proposed an unsupervised domain adaptation method for solving cross-machine fault diagnosis problems that uses Wasserstein distance to measure the similarity between samples, determine soft pseudo-labels, and improve the accuracy of fault diagnosis in the target domain.

The aforementioned methods all employ a single-task training approach, which is prone to the problem of data specification. Regarding the large differences between different machines, exploring universal fault features is more helpful for solving cross-machine fault diagnosis problems. Multi-task learning provides a powerful tool that can enhance the model's adaptive learning ability by focusing on the common knowledge between multiple related tasks [20]. Multi-task learning can uncover the underlying general features among different tasks and alleviate overfitting by adjusting the model's attention to multiple tasks, which has a positive guiding effect on the completion of transfer diagnosis tasks. Xie et al. [21] proposed a novel multi-task attention-guided network for achieving multi-objective fault diagnosis under small samples. The method consists of a task-shared network capable of learning global features and a task-customized attention network capable of completing different tasks, enabling the simultaneous training of different tasks and the extraction of common domain knowledge valuable for multiple tasks. Moreover,

to improve the model's task adaptability, an adaptive weighting method was proposed to adjust the weight parameters of the multi-task model. Kong et al. [22] proposed a multi-task self-supervised method to obtain diagnostic knowledge from unlabeled data, designing three self-supervised learning tasks to extract information contained in vibration signals from different levels, completing fault diagnosis tasks for bearings and motors. In summary, multi-task learning can learn general fault diagnosis knowledge and has certain advantages in fault diagnosis. However, the aforementioned methods all have attached some constraints to multi-task learning design, without paying attention to the temporal correlation of vibration data. Therefore, introducing multi-scale information of vibration data, three multi-scale self-supervised learning tasks are proposed.

Combining the advantages of transfer learning and multi-task learning may provide assistance in addressing the issue of cross-machine rolling bearing fault diagnosis. Therefore, the multi-task self-supervised transfer learning framework (MTSTLF) is proposed for rolling bearing fault diagnosis on cross-machine. The main contributions are summarized as follows:

1. A novel multi-task self-supervised transfer learning framework is proposed for rolling bearing fault diagnosis on cross-machine. The proposed method is trained on three self-supervised tasks and one fault diagnosis task, which can extract general transferable fault diagnosis features via multi-task learning. By multi-task learning, the ability of the model to resist overfitting has been improved. The effectiveness of the proposed method is evaluated on three datasets from different test-beds. The experimental results indicate that the proposed multi-task framework can complete cross-machine rolling bearing fault diagnosis effectively.
2. Multi-scale self-supervised learning tasks are constructed to fully extract the intrinsic information from the original vibration data. Leveraging the label-free data feature extraction capability of self-supervised learning, the model is more capable of focusing on the data rather than the task level. Moreover, different self-supervised training tasks focus on different intrinsic features of the data, which enhances the diversity in the model's feature extraction capabilities.
3. A multi-perspective feature transfer method based on Wasserstein distance and cosine similarity is proposed to acquire transferable fault diagnosis knowledge, thereby completing fault diagnosis of rolling bearings on cross-machines. By integrating probability distribution metrics and geometric similarity metrics, the model pays attention to diverse transferable diagnostic knowledge from different levels, enhancing the model's transfer diagnostic ability, and thus efficiently accomplishing the task of cross-machine bearing fault diagnosis.

The rest of this paper is organized as follows: In Section 2, the related researches about self-supervised learning, transfer learning, and multi-task learning are introduced briefly. The proposed method is described in detail in Section 3. In Section 4, two cross-machine rolling bearing transfer experiments are carried out to verify the effectiveness of the proposed method. The conclusions are drawn in Section 5.

2. Related Works

2.1. Self-Supervised Learning

Self-supervised learning is a branch of unsupervised learning that trains models using unlabeled data to automatically generate labels. It enables models to learn valuable knowledge from unlabeled data by learning useful representations and parameters. Stacked autoencoders and deep belief networks can be considered early forms of self-supervised learning, which learn from unlabeled data through a layer-wise greedy training approach. After the model is pre-trained, it is usually fine-tuned. The pre-trained parameters of the model are transferred to downstream tasks, which can improve the model's performance on the target task with simple training from scratch.

The design of pre-training tasks takes various forms. They often involve transforming or masking a part of the original input data, with the aim of forcing the model to

predict the missing parts of the data, thereby learning the intrinsic features of the data. Self-supervised learning can be divided into generative and discriminative types. Generative self-supervised learning takes the entire or part of the input data as the model's predicted output, with the goal of making the model's generated output closer to the original input data [23,24]. Typical generative self-supervised learning models consist of an encoder and a decoder, such as generative adversarial networks (GANs) [25]. Discriminative self-supervised learning is trained by discriminating positive and negative samples, requiring well-designed sample pairs for training. Typical methods include contrastive learning [26,27], which learns valuable knowledge in the process of distinguishing the authenticity of samples. However, the discriminative self-supervised learning method needs manually created sample pairs, which relies on expert experience. This paper uses a generative self-supervised learning approach and does not perform pre-training, directly integrating the constructed self-supervised tasks into the proposed multi-task learning framework.

2.2. Transfer Learning

Transfer learning primarily aims to acquire transferable knowledge from source domain tasks and then apply this knowledge to related but distinct target domain tasks. Transfer learning encompasses two concepts: domain and task. A domain consists of a feature space and a probability distribution, while a task includes a label space and a decision function, which must be learned from sample data. In practice, a domain can be observed through examples with or without labels. For instance, the source domain is composed of labeled examples, whereas the target domain contains entirely unlabeled examples. Transfer learning utilizes the knowledge obtained in the source domain to improve the decision function learned in the target domain, making the target domain decision function better adapted to the target domain task. For cross-machine rolling bearing fault diagnosis, transfer learning involves training a model on the data from one machine and then transferring and applying it to the data from another machine, ensuring good diagnostic performance.

Domain adaptation is widely applied in the field of fault diagnosis. The maximum mean discrepancy (MMD) is usually chosen as a means of domain adaptation [28,29]. MMD uses the method of mapping features into a high-dimensional kernel space for computation, which is not conducive to the gradient computation process during model training [30]. The Wasserstein distance does not have the shortcomings of MMD and is therefore chosen as the measurement method.

2.3. Multi-Task Learning

Multi-task learning was initially widely applied in fields such as computer vision and natural language processing, capable of fully leveraging the intrinsic correlation information between different tasks to handle multiple tasks simultaneously. Unlike traditional deep learning, which processes each task independently, multi-task learning aims to achieve multiple task objectives by simultaneously optimizing the loss functions of multiple tasks [31]. Multi-task learning is similar to the human brain's ability to handle multiple tasks and has the following advantages: (1) Multi-task learning can avoid the repetitive learning of general features across tasks, improving training efficiency. (2) Multi-task learning can learn more generalized features by reducing the interference of noise information from multiple tasks. (3) Multi-task learning can train a generalized model by balancing the optimization objectives of multiple tasks, making the model more adaptable. (4) Multi-task learning can focus on the correlation features between multiple tasks, learning more complex knowledge and alleviating overfitting phenomena in single-task frameworks.

3. The Proposed MTSTLF Method

The proposed method consists of self-supervised learning task and transfer fault diagnosis task. As shown in Figure 1, the self-supervised learning consists of feature

encoder and feature decoder. The backbone of the self-supervised learning is convolutional and deconvolutional networks, with the encoder using a convolutional network and the decoder using a deconvolutional network. These are designed as three parallel network branches to simultaneously process three self-supervised tasks. The fault classifier utilizes three fully connected layers and incorporates Wasserstein distance and cosine similarity into the loss function. These metrics measure the differences between data from various aspects, and by leveraging the backpropagation algorithm, the parameters are updated to achieve transfer fault diagnosis. The parameters of the proposed method are as shown in Table 1, with the decoder structures of the three branches being consistent with the decoder network parameters.

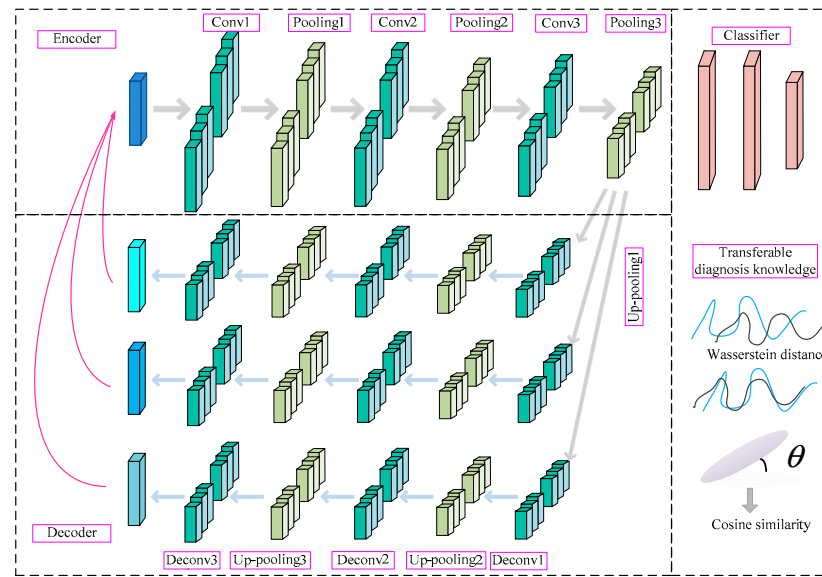


Figure 1. MTSTLF structure.

Table 1. Details on the architecture of the proposed method.

Modules	Description of Layer	Parameter
Feature encoder	Kernel size of first convolutional layer	1*64
	Number of units (input, output) in first convolutional layer	(1, 16)
	Kernel size and stride of max-pooling layer	2*2/2
	Kernel size of second convolutional layer	1*3
	Number of units in second convolutional layer	(16, 16)
	Kernel size and stride of max-pooling layer	2*2/2
	Kernel size of forth convolutional layer	1*3
	Number of units in third convolutional layer	(16, 16)
	Kernel size of third convolutional layer	1*3
Feature decoder	Kernel size and stride of max-pooling layer	2*2/2
	Kernel size of first deconvolutional layer	1*3
	Number of units (input, output) in first deconvolutional layer	(16, 16)
	Kernel size and stride of up max-pooling layer	2*2/2
	Kernel size of second deconvolutional layer	1*3
	Number of units (input, output) in second deconvolutional layer	(16, 16)
	Kernel size and stride of up max-pooling layer	2*2/2
Classifier	Kernel size of third deconvolutional layer	1*3
	Number of units (input, output) in third deconvolutional layer	(16, 1)
	Kernel size and stride of up max-pooling layer	2*2/2
Classifier	Number of units (input, output) in first full-connected layer	16*120/512
	Number of units (input, output) in second full-connected layer	512/512
	Number of units (input, output) in third full-connected layer	512/4

At present, there is not a systematic method to determine the structure of various deep learning models and the tasks or hyperparameters in the proposed method are confirmed by debugging experience and experiments.

The model is built based on the PyTorch 1.14.0 deep learning framework and Python 3.11, which runs on Windows 10 with 8 GB RAM.

3.1. Multi-Scale Self-Supervised Learning Tasks

Usually, due to the collisions between equipment components during operation, the vibration signals exhibit a multi-resonance phenomenon across the frequency range. The characteristic information of the vibration signals propagates on multiple time scales, and focusing on the information of the vibration signals at different time scales is helpful for fully extracting valuable fault diagnostic features [32]. Therefore, a multi-scale self-supervised learning task design method has been proposed for the commonly used masking method in designing self-supervised tasks.

As shown in Figure 2, three masking methods of different scales have been designed. The vibration amplitude of the signal is set to 0 at the time points where masking is required, so the process of self-supervised learning is the process of restoring the original signal based on the masked signal. In this process of self-supervised learning, the model can learn the intrinsic multi-scale characteristics and periodic information of the vibration signal. Given an input sample $x = \{x_1, x_2, \dots, x_n\}$, $n > 0$, n indicates the number of data points contained in the input sample, n can be divided by 2. The mask formulas for different time scales are defined as follows:

$$x = \{x_1, x_2, x_3, \dots, x_{n-1}, x_n\}, [x_{1+2*\tau*(i-1)}, x_{(2i-1)*\tau}] = 0, 1 \leq i \leq n/(2*\tau) + 1 \quad (1)$$

where τ is the time scale; $[\bullet]$ denote the numerical values from the sample in order, including the front and back values.

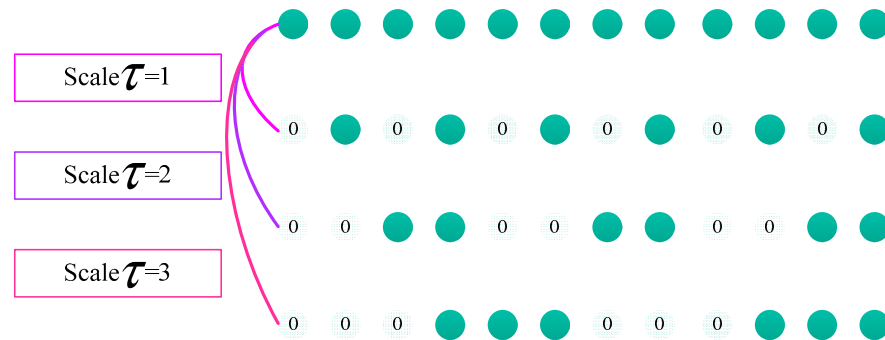


Figure 2. The procedure of three masking methods.

To realize restoring signals via self-supervised learning, the mean square error is selected as the loss function. The loss of the self-supervised process is ultimately defined as

$$\mathcal{L}_{self-supervised} = \frac{1}{3n} \sum_{\tau=1}^3 \sum_{i=1}^n (y_{(i)} - x_{(i)})^2 \quad (2)$$

where \mathcal{L} denotes the final loss of three self-supervised tasks, and y is the output of the model with the input x .

3.2. Multi-Perspective Feature Transfer Method

Maximum mean discrepancy (MMD) is one of the typical measures of probability distributions, often used in transfer learning to calculate the distribution differences between different domain data, serving as an optimization target to train models and narrow the distribution of cross-domain data. MMD uses the method of mapping features into

a high-dimensional kernel space for computation, which is not conducive to the gradient computation process during model training. The Wasserstein distance does not have the shortcomings of MMD and is therefore chosen as the measurement method. The Wasserstein distance is a measure commonly used to explore the shortest distance or minimum difference between two distributions, and it is very important in the comparison of probability distributions. The definition of the Wasserstein distance is as follows:

$$W(P(X), P(Y)) = \inf_{\mu \in \Pi(P(X), P(Y))} \mathbb{E}_{(x,y) \sim \mu} [\|x - y\|] \quad (3)$$

where $\Pi(P(X), P(Y))$ denotes the joint distribution $\mu(x, y)$ of (x, y) , $P(X)$ is the margin distribution of variable x , $P(Y)$ is the margin distribution of variable y , and $\inf(\bullet)$ denotes infimum.

Cosine similarity is the cosine of the angle between two vectors in space, reflecting the relative spatial position of the vectors, and is a geometric measure of their orientation. Given two random variables X and Y , the cosine similarity between them is calculated as follows:

$$\text{Cosine}(X, Y) = \frac{X \bullet Y}{\|X\| * \|Y\|} = \frac{\sum_{i=1}^n X_i * Y_i}{\sqrt{\sum_{i=1}^n X_i^2} * \sqrt{\sum_{i=1}^n Y_i^2}} \quad (4)$$

Therefore, by narrowing the distribution differences between the health condition features of cross-machine bearings from the perspective of probability distribution, the model can learn features with similar probability distributions. From the perspective of geometric similarity, by reducing the spatial position angles between the health condition features of cross-machine bearings, the model can extract positional information between features. Integrating probability distribution and geometric similarity allows for a multi-faceted observation and learning of transferable fault diagnosis knowledge, helping the model to extract rich transferable diagnostic knowledge, thereby enhancing the model's transfer fault diagnosis capability. The loss of multi-perspective feature transfer method is defined as

$$\mathcal{L}_{trans} = W(P(S), P(T)) + \text{Cosine}(X, Y) \quad (5)$$

3.3. Multi-Task Learning of the Proposed Method

The multi-task learning of the proposed method includes data reconstruction and fault diagnosis. In data reconstruction task, MSE is selected as loss function. In fault diagnosis task, cross-entropy is selected as loss function and the Softmax is the classifier. Furthermore, the loss of multi-view confusion transfer is designed to realize transfer fault diagnosis. In multi-task learning, different loss functions are designed. During training, the weights of the model will not be biased by the task represented by one of the loss functions, and all tasks will affect the training of other tasks. Therefore, the final model obtained is trained through balancing multiple tasks, which ensures that the data do not exhibit rapid convergence or even overfitting on a single task, thus alleviating overfitting. So, the loss of multi-task learning can be calculated as follows:

$$\mathcal{L}_{multi-task} = \mathcal{L}_{self-supervised} + \mathcal{L}_{trans} + \mathcal{L}_{diagnosis} \quad (6)$$

3.4. The Procedure for the Proposed Method

The procedure for the MTSTLF method, as presented in Figure 3, mainly includes three stages: data acquisition, multi-task learning, and fault diagnosis. The general procedures of the proposed method are summarized as follows:

Step 1: Use a data acquisition system containing test rig and acceleration transducer to collect various health condition vibration signals of rolling bearings on different machines,

divided into a source domain dataset (training dataset) and a target domain dataset (testing dataset). All data collected from laboratory environment.

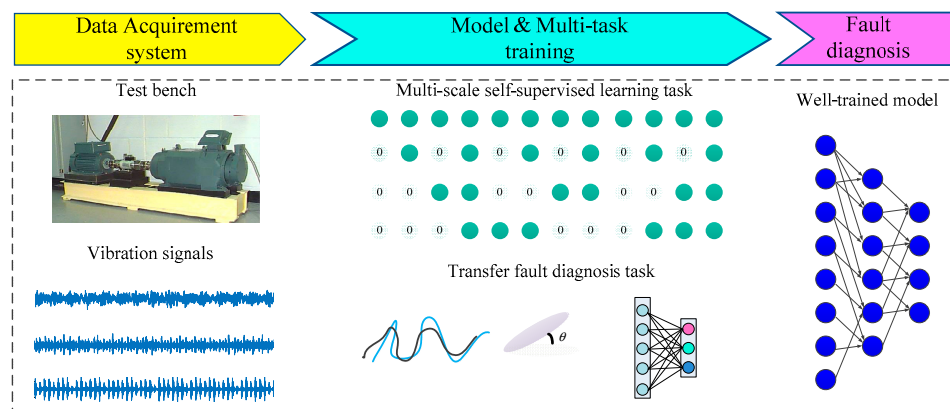


Figure 3. Fault diagnosis process of the proposed method.

Step 2: Input both the source domain dataset and the target domain dataset with raw vibration data into the constructed multi-task learning model simultaneously. The feature encoder extracts universal features, and the feature decoder completes self-supervised tasks based on feature extracted by the encoder. The fault classifier identifies the health status of the bearings by utilizing multiple perspectives information based on Wasserstein distance and cosine similarity to extract transferable fault diagnosis knowledge.

Step 3: Use the well-trained model to complete fault diagnosis on the target domain dataset.

4. Experiment Verification

4.1. Experiment 1: CWRU and Ottawa

4.1.1. Dataset Description

The widely used benchmark bearing fault data are supplied through Case Western Reserve University (CWRU) [33]. As shown in Figure 4, the test setup mainly consists of a motor, a torque transducer/encoder, and a dynamometer. Single-point faults are introduced on rolling bearings by using the electro-discharge machining technique. In this paper, the data utilized are from the drive end bearing under 0hp, corresponding to a motor speed of 1797 rpm. The data are acquired by an accelerometer transducer, which is fixed at the top of the bearing block with a sampling frequency of 12 kHz. The experimental dataset includes three fault types with fault diameters of 0.007 inches, i.e., normal condition (N), inner race fault (IF), ball fault (BF), and outer race fault (OF). Each health condition contains 40 samples, and each sample has 1024 data points. The signal waveform is partially presented in Figure 5b.

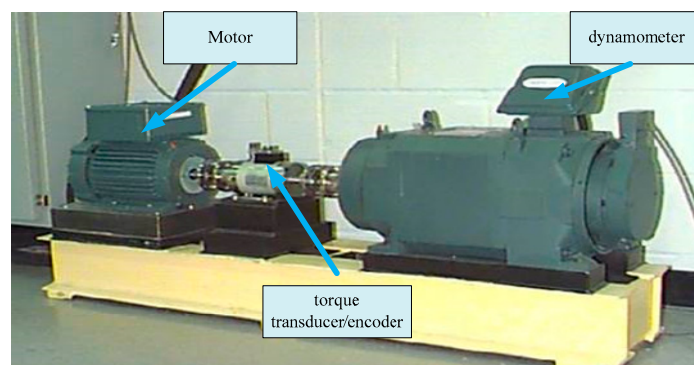


Figure 4. Experimental setup of CWRU.

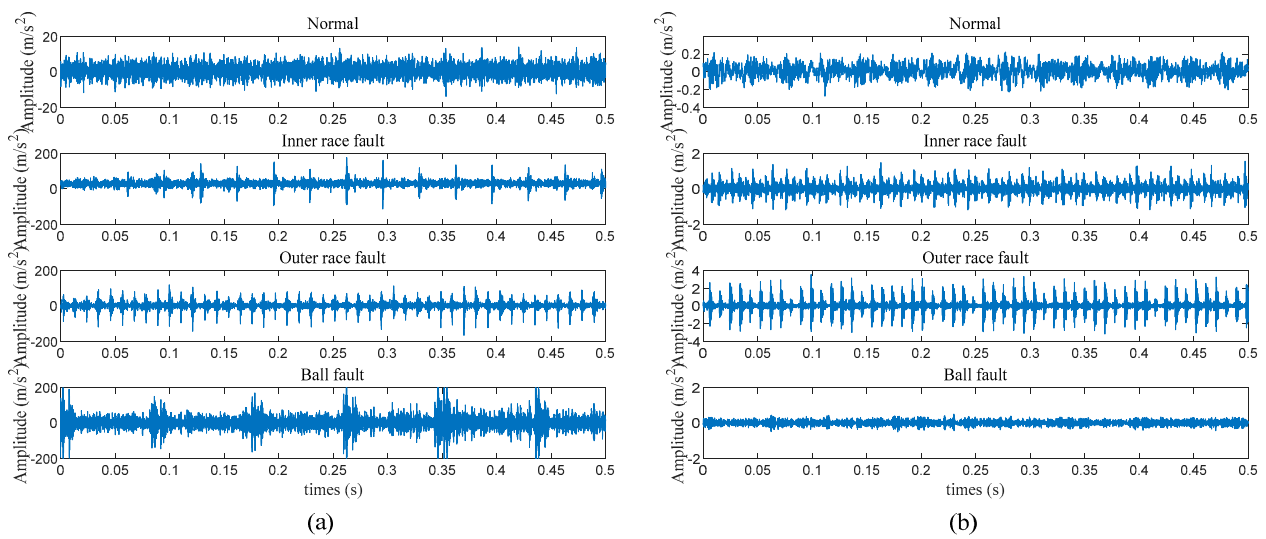


Figure 5. The waveform of raw vibration signals: (a) Ottawa; (b) CWRU.

Another dataset is from the University of Ottawa [34]. The experimental test rig consists of a single-phase motor mounted on a rigid plate supported by anti-vibration mounts as shown in Figure 6. The shaft is stepped up using a shaft adapter, on which an SKF E22206 spherical roller bearing is mounted to withstand the load applied by the cantilever beam. The motor is driven at a constant nominal speed of 1750 rpm and a load of 400 N. The sampling frequency of the data is 42 kHz, and 10 s of data are collected for each health condition. The health status of the bearings used in this experiment is healthy, with ball, inner, and outer race failures. Each health status contains 40 samples, and the sample length of each sample is 1024. The original signal waveforms of different health conditions are shown in Figure 5a.

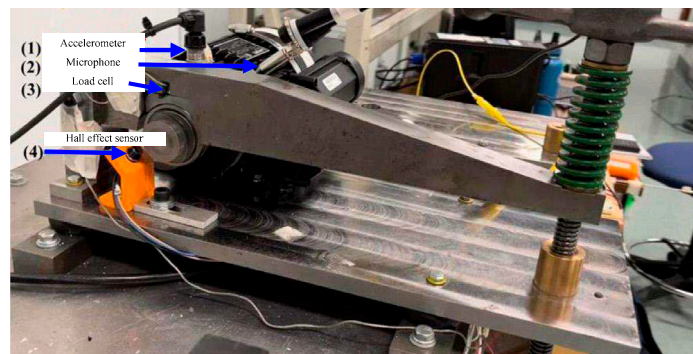


Figure 6. Experimental setup of Ottawa.

Two cross-machine transfer fault diagnosis experiments were designed using the above two datasets. The specific configuration of the experiment is shown in Table 2.

Table 2. The detailed experimental description of cross-machine rolling bearing fault diagnosis.

Task Number	Source Domain	Target Domain	Fault Types and Label	The Number of Samples from Each Health Condition
T0	CWRU	Ottawa	Normal/1 Inner race fault/2 Outer race fault/3 Ball fault/4	40
T1	Ottawa	CWRU	Normal/1 Inner race fault/2 Outer race fault/3 Ball fault/4	40

4.1.2. Experimental Result Analysis

To compare and verify the effectiveness of the proposed method, Transfer Component Analysis (TCA), Deep Domain Confusion (DDC), a Deep Adaptation Network (DAN), and Domain Consensus Clustering (DCC) [35] were implemented on the cross-machine rolling bearing fault dataset. The network structure and parameters are stated as follows (it is worth noting that all models use raw vibration data as input):

4. TCA is a classic transfer learning method that uses MMD as the measurement method to map the cross-domain data to the same space for distribution difference calculation without combining in-depth learning. Through grid search technology, the weight coefficient of TCA regularization term is set to 0.7, and the subspace dimension is 12.
5. The basic structure of DDC is consistent with the proposed method, with the decoder removed and only MMD used as the migration distance metric.
6. The basic structure of the DAN is consistent with the proposed method, which removes the decoder. The DAN uses multi-kernel MMD as the transfer distance metric and calculates the distance of multi-layer output features.
7. The design of the DCC method is consistent with the original literature.

The epoch of DDC, DAN, DCC, and the proposed method is 1000. The learning rate of DDC and proposed method is 0.0001 and the learning rate of DAN is 0.001.

The comparative experimental results of different methods are presented in Table 3. From Table 3, it can be observed that the proposed method achieves higher fault diagnosis accuracy on both transfer tasks T0 and T1 compared to the other four methods, indicating that the proposed method possesses strong cross-machine migration fault diagnosis capabilities. A comparison of the experimental results between the DAN and the proposed method leads to the conclusion that multi-task learning can enhance the ability to extract valuable features. The knowledge acquired by the model is universal and is not influenced by the customized weights of the model trained on a single task. In contrast to the single approach of achieving transfer fault diagnosis by narrowing the probability distribution, the multi-perspective feature transfer method, which integrates geometric similarity, has a more powerful transfer learning capability, thereby improving the model's precision in cross-machine rolling bearing fault diagnosis tasks. Comparing the two transfer learning methods, DAN and DDC, it can be seen that the DAN has a higher transfer accuracy rate than DDC. The reason is that the DAN focuses on richer transferable knowledge due to using multi-kernel MMD to reduce the distribution difference between the source and target domain data and calculate the distribution distance of multiple feature layers. Due to TCA only using MMD to reduce the distribution difference between the source and target domain data, and extracting only shallow features, it lacks the adaptive learning ability of deep learning, resulting in the lowest accuracy.

Table 3. The results of different contrastive methods.

Method	Accuracy of T0 (%)	Accuracy of T1 (%)
TCA	53.13	51.25
DDC	59.38	55.63
DAN	69.38	62.50
DCC	80.00	81.25
MTSTLF (ours)	91.87	91.25

In order to observe the recognition ability of five methods for different health conditions, a confusion matrix was used to demonstrate the effectiveness, as shown in Figure 7. Figure 7e shows the recognition results of the proposed method. From the numbers on the diagonal, it can be seen that the recognition accuracy for different health conditions is 100%. The results indicate the following: (1) The proposed method can effectively extract transferable knowledge and can be fully applied to the data of new machines. (2) The proposed method can extract distinguishable features of different health conditions, effectively identifying different health conditions of rolling bearings from the data. Compared

to other methods that can accurately identify normal and inner circle faults, the recognition performance of TCA is slightly inferior, indicating its poor ability to extract distinguishable transfer features.

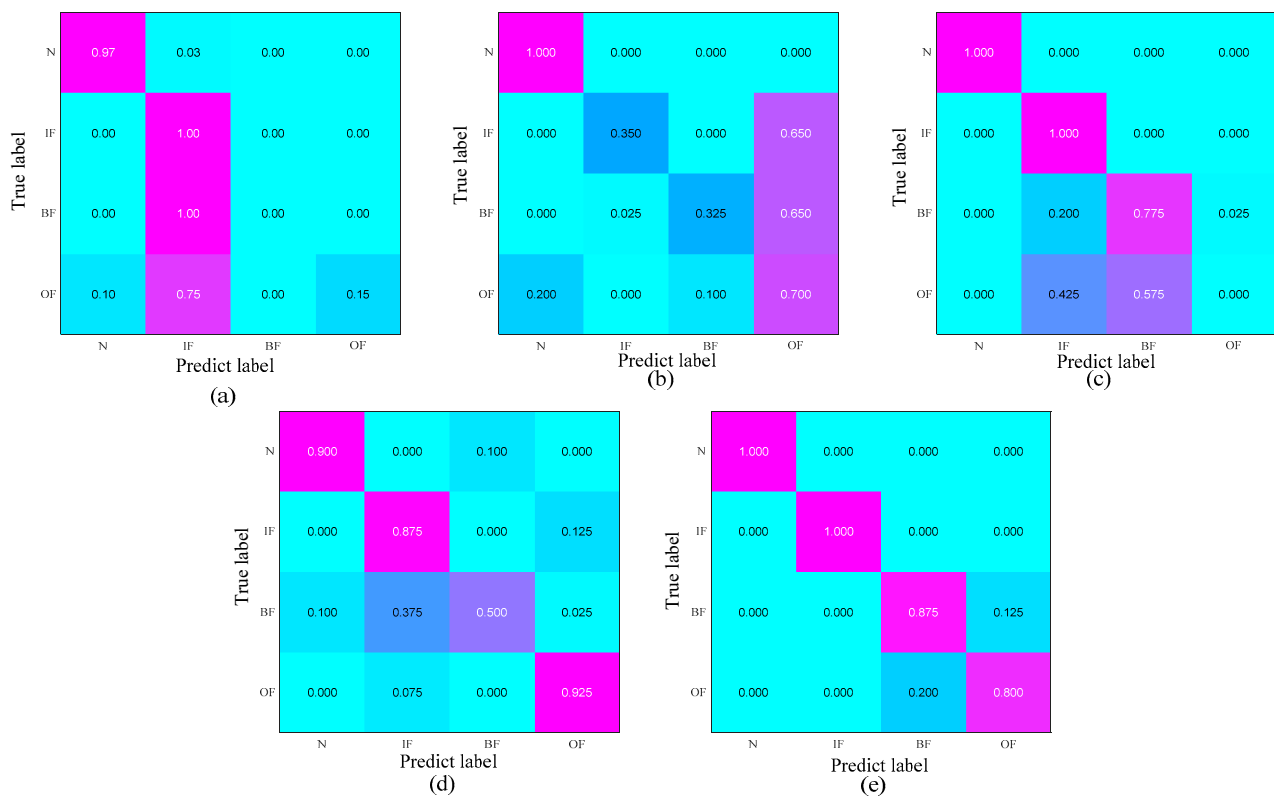


Figure 7. The confusion matrix of experimental results: (a) TCA; (b) DDC; (c) DAN; (d) DCC; (e) MTSTLF (ours).

To clearly observe the transfer fault diagnosis effect of cross-machine rolling bearings, t-distributed stochastic neighbor embedding (t-SNE) [36] was used to visualize the outputs of several methods, as shown in Figure 8. From Figure 8a, it can be seen that only the inner race fault is accurately distinguished, and the characteristics of normal and outer race faults are difficult to distinguish when they are mixed together. Comparing Figure 8a–e, the separability of various health conditions in Figure 8e is the best, and the distance between different health conditions is relatively large. The characteristics of the same health condition are closely concentrated together, indicating that the proposed MTSTLF method has the best classification effect and accurately identifies various health conditions of cross-machine rolling bearings. From Figure 8d, it can be seen that a few normal features and ball fault features are confused. This makes the model prone to misjudgment when distinguishing between the two health conditions, and the cross-machine transfer fault diagnosis ability is low. From Figure 8e, the features between BF and OF are partially overlapping, indicating that the proposed method is prone to misjudging above types of faults. From 8a–e, the features from normal condition are distinguishable. The reason is that the representative characteristics of damaged bearings and normal bearings are completely different.

In order to observe the performance of different comparison methods in terms of computational burn, the time used in the training and testing phases at each epoch was counted. As shown in Table 4, the training time and testing time of the proposed method are not the worst. Compared with the proposed method, DCC achieved an improvement in diagnostic accuracy of about 10% in no more than three times the DCC time. Taking into account both computational burn and diagnostic accuracy, the proposed method does not

differ significantly in terms of time, but does have a significant improvement in accuracy and certain advantages.

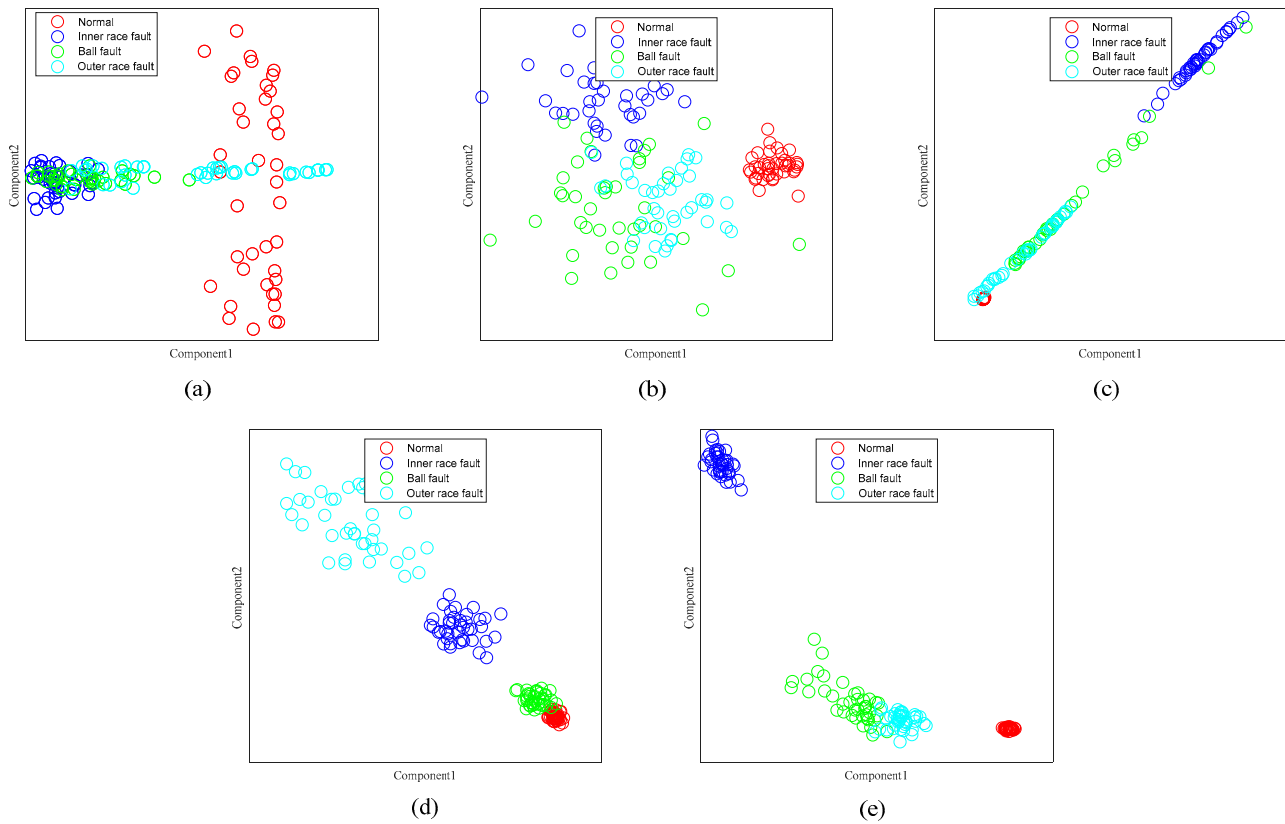


Figure 8. The clustering visualization of experimental results: (a) TCA; (b) DDC; (c) DAN; (d) DCC; (e) MTSTLF (ours).

Table 4. The computational time of different contrastive methods.

Method	Training Time (s)	Testing Time (s)
TCA	/	2.6
DDC	1.80	0.38
DAN	7.35	0.43
DCC	2.70	0.88
MTSTLF (ours)	6.48	0.83

4.2. Experiment 2: CWRU and SEU

4.2.1. Data Description

The bearing dataset from CWRU is described in Section 4.1.1 and will not be further elaborated here.

The other dataset uses a gearbox dataset from Southeast University, collected from the Drivetrain Dynamic Simulator [37]. The experimental platform consists of a motor controller, a motor, a planetary gearbox, a reduction gearbox, a load, and a load controller. This experiment used a subset of bearing fault data collected from the operational condition with a speed load configuration of 20 Hz–0 V (0 Nm) and a sampling frequency of 5120 Hz. The health conditions of the experiment using bearing data are normal, ball failure, inner race failure, and outer race failure, with each health condition containing 40 samples and a sample length of 1024 for each sample. The transfer tasks are listed in Table 5.

Table 5. The detailed experimental description of cross-machine rolling bearing fault diagnosis.

Task Number	Source Domain	Target Domain	Fault Types and Label	The Number of Samples from Each Health Condition
T2	CWRU	SEU	Normal/1 Inner race fault/2 Outer race fault/3 Ball fault/4	40
T3	SEU	CWRU	Normal/1 Inner race fault/2 Outer race fault/3 Ball fault/4	40

4.2.2. Experimental Result Analysis

For comparing and verifying the effectiveness of the proposed method, Transfer Component Analysis (TCA), Deep Domain Confusion, a Deep Adaptation Network, and Domain Consensus Clustering (DCC) were implemented. The network structure and parameters are same as in Section 4.1.2.

The comparative experimental results of different methods are shown in Table 6 and Figure 9. From Table 6, it can be seen that the proposed method MTSTLF (ours) has an accuracy of 98.33% and 96.67% on two cross-machine migration fault diagnosis tasks, T2 and T3, respectively, which are higher than the recognition accuracy of the other four methods. This indicates that the proposed method has a relatively strong cross-machine transfer fault diagnosis ability. Comparing the experimental results of DDC and DAN methods, it can be seen that the accuracy of the DAN is significantly higher than that of DDC. This is because the DAN uses multi-kernel MMD to reduce the distribution difference between the source and target domain data, and calculates the distribution distance of multiple feature layers, which focuses on richer transferable knowledge and improves the transfer diagnosis ability. As shown in Figure 9, the cross-machine transfer fault diagnosis accuracy of TCA is the lowest, mainly because the features extracted by TCA are relatively shallow and do not have the adaptive learning ability of deep learning, and the internal attention to the data is not deep enough.

Table 6. The results of different contrastive methods.

Method	Accuracy of T2 (%)	Accuracy of T3 (%)
TCA	56.25	53.75
DDC	71.25	66.25
DAN	74.37	77.50
DCC	83.13	83.75
MTSTLF (ours)	95.63	90.00

In order to clearly observe the migration fault diagnosis effect of cross-machine rolling bearings, t-SNE was used to visualize the outputs of the five methods mentioned above, as shown in Figure 10. From Figure 10a, it can be seen that only normal health conditions are accurately distinguished, and the characteristics of inner and outer race faults are mixed together and difficult to distinguish. Comparing Figure 10a–e, the separability of various health conditions in Figure 10e is the best, and the distance between different health conditions is relatively large. The characteristics of the same health condition are closely concentrated together, indicating that the proposed MTSTLF method has the best classification effect and accurately identifies various health conditions of cross-machine rolling bearings. From Figure 10d, it can be seen that although most of the features of the inner and outer race faults are concentrated together, there are still a few outer race features scattered around the inner race fault features. This makes the model prone to misjudgment when distinguishing between the two health conditions, and the cross-machine transfer

fault diagnosis ability is low. By comparing the experimental results in Figure 10a–e with those in Section 4.1.2, it is not difficult to find that among the three health conditions, normal health characteristics are the easiest to distinguish and have stronger distinguishability.

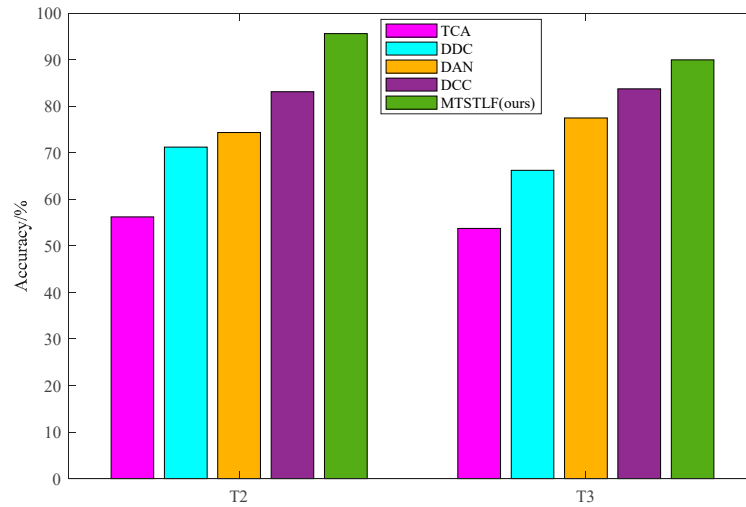


Figure 9. The results of different classification methods on T2 and T3.

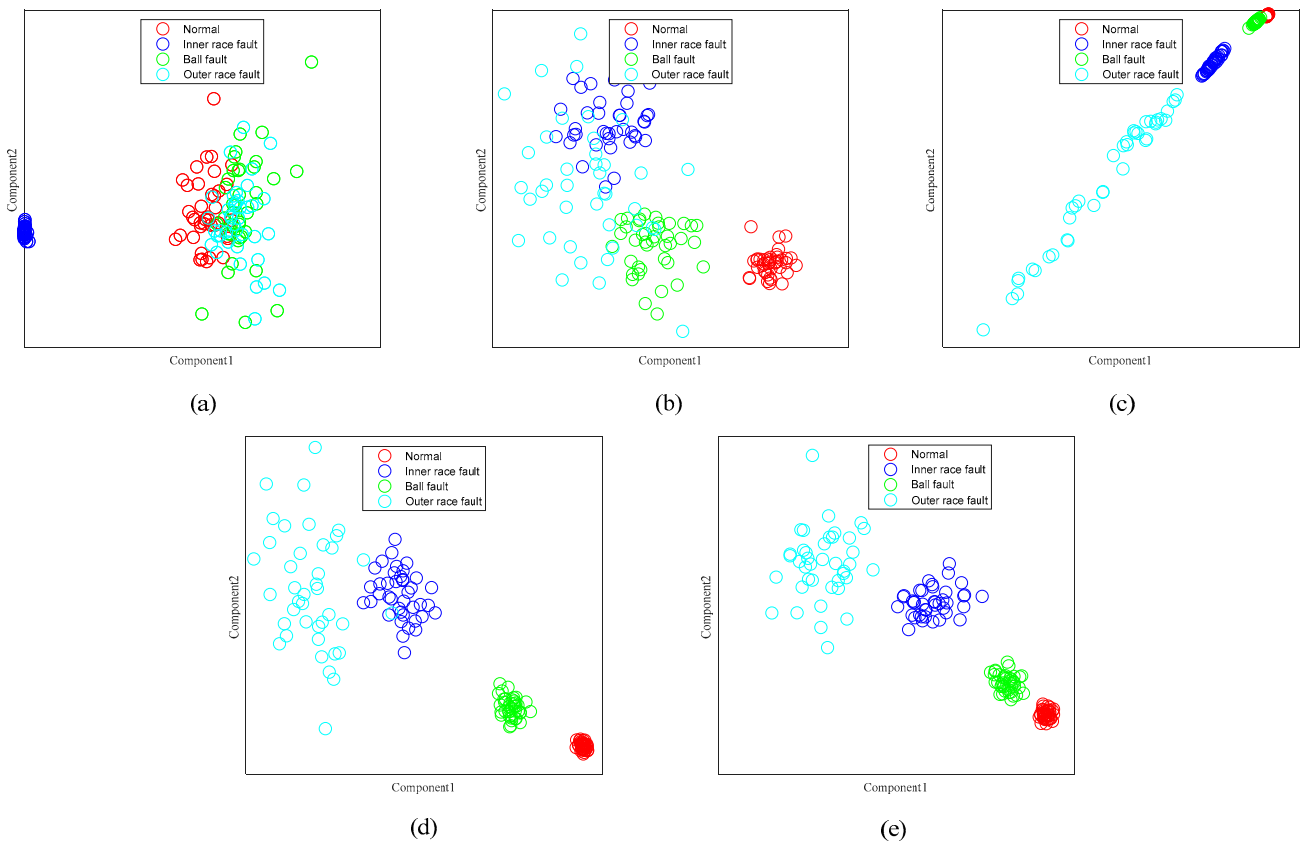


Figure 10. Visualization results of clustering using different methods: (a) TCA; (b) DDC; (c) DAN; (d) DCC; (e) MTSTLF (ours).

Taking into account tasks T0, T1, T2, and T3, the average accuracy of TCA is the worst, and the average accuracy of the proposed method is the best. To some extent, multiple tasks reflect the average performance of the proposed method, which has more advantages compared to other migration methods.

Comparing the results in Tables 3 and 5, it can be seen that the accuracy of CWRU/SEU is better than that of CWRU/Ottawa on the whole regarding the different methods. Therefore, the greater the difference in operating conditions, the worse the migration effect.

5. Conclusions

A novel multi-task self-supervised transfer learning framework (MTSTLF) is proposed for cross-machine rolling bearing fault diagnosis. The proposed method is trained using a multi-task learning paradigm, consisting of self-supervised learning tasks and fault diagnosis tasks. Based on the multi-scale information contained in the vibration signals of rolling bearings, three different data masking methods are designed. By training the proposed multi-task model through self-supervised learning, the attention paid to the intrinsic features of data in different health conditions is enhanced, thereby strengthening the ability of the feature extractor to extract task-universal features. A multi-perspective feature transfer method is proposed for completing cross-machine fault diagnosis tasks. Focusing on probability distribution and geometric similarity, the method learns fault diagnosis knowledge that is transferable and has strong discriminative power, thereby enhancing the transfer learning ability and improving the adaptability of the proposed method to complex working conditions. Two experimental cases are carried out to evaluate the effectiveness of the proposed method. The experimental results demonstrate that the proposed method has strong transfer fault diagnosis capabilities. It can mitigate the negative impact of machine-specific attributes on the diagnostic ability and can complete cross-machine fault diagnosis tasks with high precision.

In forthcoming research, we will analyze and discuss more challenges in adapting models across different machines, such as machine-specific noise and sensor variability.

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

Funding: This research was funded by National Natural Science Foundation of China (52177042), the Hebei Provincial Natural Science Foundation (E2022502003), the Chinese Fundamental Research Funds for the Central Universities (2023MS128), the Top Youth Talent Support Program of Hebei Province ([2018]-27), the Hebei Provincial High-Level Talent Funding Project (B20231006), and the South Taihu Innovative Talent Team Project.

Data Availability Statement: The original data presented in the study are openly available in CWRU (Case Western Reserve University) Bearing Dataset at <https://engineering.case.edu/bearingdatacenter> (accessed on 14 November 2024).

Conflicts of Interest: Authors Honghua Bai and Weiling Huang were employed by the company Zhejiang Zhenxing Axiang Group. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ma, S.; Chu, F.L. Ensemble deep learning-based fault diagnosis of rotor bearing systems. *Comput. Ind.* **2019**, *105*, 143–152. [[CrossRef](#)]
2. Wu, C.M.; Zheng, S.P. Fault diagnosis method of rolling bearing based on MSCNN-LSTM. *Comput. Mater. Contin.* **2024**, *79*, 4395–4411. [[CrossRef](#)]
3. Liu, G.Z.; Wu, L.F. Incremental bearing fault diagnosis method under imbalanced sample conditions. *Comput. Ind. Eng.* **2024**, *192*, 110203. [[CrossRef](#)]
4. Dai, X.; Yi, K.; Wang, F.L.; Cai, C.; Tang, W. Bearing fault diagnosis based on POA-VMD with GADF-Swin Transformer transfer learning network. *Measurement* **2024**, *238*, 115328. [[CrossRef](#)]
5. Lei, Y.G.; Yang, B.; Jiang, X.W.; Jia, F.; Li, N.; Nandi, A.K. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [[CrossRef](#)]

6. Zheng, B.; Huang, J.H.; Ma, X.; Zhang, X.; Zhang, Q. An unsupervised transfer learning method based on SOCNN and FBNN and its application on bearing fault diagnosis. *Mech. Syst. Signal Process.* **2024**, *208*, 111047. [[CrossRef](#)]
7. Kumar, P.; Raouf, I.; Kim, H.S. Transfer learning for servomotor bearing fault detection in the industrial robot. *Adv. Eng. Softw.* **2024**, *194*, 103672. [[CrossRef](#)]
8. Vashishtha, G.; Kumar, R. Unsupervised Learning Model of Sparse Filtering Enhanced Using Wasserstein Distance for Intelligent Fault Diagnosis. *J. Vib. Eng. Technol.* **2023**, *11*, 2985–3002. [[CrossRef](#)]
9. Wang, R.; Yan, F.C.; Yu, L.; Shen, C.; Hu, X. Joint Wasserstein distance matching under conditional probability distribution for cross-domain fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.* **2024**, *210*, 111121. [[CrossRef](#)]
10. Zhao, K.; Jia, F.; Shao, H. A novel conditional weighting transfer Wasserstein auto-encoder for rolling bearing fault diagnosis with multi-source domains. *Knowl. Based Syst.* **2023**, *262*, 110203. [[CrossRef](#)]
11. Islam, M.; Zunair, H.; Mohammed, N. CosSIF: Cosine similarity-based image filtering to overcome low inter-class variation in synthetic medical image datasets. *Comput. Biol. Med.* **2024**, *172*, 108317. [[CrossRef](#)] [[PubMed](#)]
12. Gao, C.; Li, W.F.; He, L.J.; Zhong, L. A distance and cosine similarity-based fitness evaluation mechanism for large-scale many-objective optimization. *Eng. Appl. Artif. Intell.* **2024**, *133 Pt B*, 108127. [[CrossRef](#)]
13. Akram, F.; Ahmad, T.; Sadiq, M. An integrated fuzzy adjusted cosine similarity and TOPSIS based recommendation system for information system requirements selection. *Decis. Anal. J.* **2024**, *11*, 100443. [[CrossRef](#)]
14. Lu, F.Y.; Tong, Q.B.; Jiang, X.D.; Feng, Z.; Xu, J.; Wang, X.; Huo, J. A deep targeted transfer network with clustering pseudo-label learning for fault diagnosis across different machines. *Mech. Syst. Signal Process.* **2024**, *213*, 111344. [[CrossRef](#)]
15. Yang, B.; Lei, Y.; Jia, F.; Xing, S. A transfer learning method for intelligent fault diagnosis from laboratory machines to real-case machines. In Proceedings of the International Conference on Sensing, Diagnostics, Prognostics, and Control, Xi'an, China, 15–17 August 2018; pp. 35–40.
16. Yang, B.; Lei, Y.; Jia, F.; Xing, S. An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings. *Mech. Syst. Signal Process.* **2019**, *122*, 692–706. [[CrossRef](#)]
17. Guo, L.; Lei, Y.; Xing, S.; Yan, T.; Li, N. Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Trans. Ind. Electron.* **2019**, *66*, 7316–7325. [[CrossRef](#)]
18. Jia, S.; Li, Y.; Wang, X.; Sun, D.; Deng, Z. Deep causal factorization network: A novel domain generalization method for cross-machine bearing fault diagnosis. *Mech. Syst. Signal Process.* **2023**, *192*, 110228. [[CrossRef](#)]
19. Jia, N.; Huang, W.G.; Ding, C.C.; Wang, J.; Zhu, Z. Physics-informed unsupervised domain adaptation framework for cross-machine bearing fault diagnosis. *Adv. Eng. Inform.* **2024**, *62 Pt C*, 102774. [[CrossRef](#)]
20. Ding, P.; Xu, Y.; Sun, X.M. Multitask learning for aero-engine bearing fault diagnosis with limited data. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 3520111. [[CrossRef](#)]
21. Xie, Z.L.; Chen, J.L.; Feng, Y.; Zhang, K.; Zhou, Z. End to end multi-task learning with attention for multi-objective fault diagnosis under small sample. *J. Manuf. Syst.* **2022**, *62*, 301–316. [[CrossRef](#)]
22. Kong, D.P.; Huang, W.D.; Zhao, L.B.; Ding, J.; Wu, H.; Yang, G. Mining knowledge from unlabeled data for fault diagnosis: A multi-task self-supervised approach. *Mech. Syst. Signal Process.* **2024**, *211*, 111189. [[CrossRef](#)]
23. Sun, X.C.; Wang, Z.H.; Lu, Z.M.; Lu, Z. Self-supervised graph representations with generative adversarial learning. *Neurocomputing* **2024**, *592*, 127786. [[CrossRef](#)]
24. Li, J.L.; Li, R.N.; Xu, L.H.; Liu, J. Self-supervised generative adversarial learning with conditional cyclical constraints towards missing traffic data imputation. *Knowl. Based Syst.* **2024**, *284*, 111233. [[CrossRef](#)]
25. Zhang, X.Y.; Shi, H.C.; Zhu, X.B.; Li, P. Active semi-supervised learning based on self-expressive correlation with generative adversarial networks. *Neurocomputing* **2019**, *345*, 103–113. [[CrossRef](#)]
26. Cao, R.L.; Li, Z.X.; Tang, Z.J.; Zhang, C.; Ma, H. Enhancing robust VQA via contrastive and self-supervised learning. *Pattern Recognit.* **2025**, *159*, 111129. [[CrossRef](#)]
27. Yang, M.; Ling, J.; Chen, J.M.; Feng, M.; Yang, J. Discriminative semi-supervised learning via deep and dictionary representation for image classification. *Pattern Recognit.* **2023**, *140*, 109521. [[CrossRef](#)]
28. Li, J.M.; Ye, Z.D.; Gao, J.; Meng, Z.; Tong, K.; Yu, S. Fault transfer diagnosis of rolling bearings across different devices via multi-domain information fusion and multi-kernel maximum mean discrepancy. *Appl. Soft Comput.* **2024**, *159*, 111620. [[CrossRef](#)]
29. Jia, S.Y.; Deng, Y.F.; Lv, J.; Du, S.; Xie, Z. Joint distribution adaptation with diverse feature aggregation: A new transfer learning framework for bearing diagnosis across different machines. *Measurement* **2022**, *187*, 110332. [[CrossRef](#)]
30. Qian, Q.; Wang, Y.; Zhang, T.S.; Qin, Y. Maximum mean square discrepancy: A new discrepancy representation metric for mechanical fault transfer diagnosis. *Knowl. Based Syst.* **2023**, *276*, 110748. [[CrossRef](#)]
31. Wang, P.C.; Xiong, H.; He, H.X. Bearing fault diagnosis under various conditions using an incremental learning-based multi-task shared classifier. *Knowl. Based Syst.* **2023**, *266*, 110395. [[CrossRef](#)]
32. Yu, J.B.; Yan, X.F. Multiscale intelligent fault detection system based on agglomerative hierarchical clustering using stacked denoising autoencoder with temporal information. *Appl. Soft Comput.* **2020**, *95*, 106525. [[CrossRef](#)]
33. Yan, X.A.; Liu, Y.; Jia, M.P. Multiscale cascading deep belief network for fault identification of rotating machinery under various working conditions. *Knowl. Based Syst.* **2020**, *193*, 105484. [[CrossRef](#)]
34. Sehri, M.; Dumond, P.; Bouchard, M. University of Ottawa constant load and speed rolling-element bearing vibration and acoustic fault signature datasets. *Data Brief* **2023**, *49*, 109327. [[CrossRef](#)] [[PubMed](#)]

35. Li, G.; Kang, G.; Zhu, Y.; Wei, Y.; Yang, Y. Domain consensus clustering for universal domain adaptation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9752–9761.
36. Li, B.T.; Pi, D.C.; Lin, Y.X. Learning ladder neural networks for semi-supervised node classification in social network. *Expert Syst. Appl.* **2021**, *165*, 113957. [[CrossRef](#)]
37. Shao, S.; McAleer, S.; Yan, R.Q.; Baldi, P. Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Trans. Ind. Inform.* **2019**, *15*, 2446–2455. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.