sensors

MDPI

Article

Multi-Scale Capsule Attention Network and Joint Distributed Optimal Transport for Bearing Fault Diagnosis under Different Working Loads

Zihao Sun [1], Xianfeng Yuan [1,*], Xu Fu [1], Fengyu Zhou [2] and Chengjin Zhang [1]

[1] School of Mechanical Electrical and Information Engineering, Shandong University, Weihai 264209, China; sunzihao@mail.sdu.edu.cn (Z.S.); fuxu@mail.sdu.edu.cn (X.F.); cjzhang@sdu.edu.cn (C.Z.)
[2] School of Control Science and Engineering, Shandong University, Jinan 250100, China; zhoufengyu@sdu.edu.cn
* Correspondence: yuanxianfeng@sdu.edu.cn

Abstract: In recent years, intelligent fault diagnosis methods based on deep learning have developed rapidly. However, most of the existing work performs well under the assumption that training and testing samples are collected from the same distribution, and the performance drops sharply when the data distribution changes. For rolling bearings, the data distribution will change when the load and speed change. In this article, to improve fault diagnosis accuracy and anti-noise ability under different working loads, a transfer learning method based on multi-scale capsule attention network and joint distributed optimal transport (MSCAN-JDOT) is proposed for bearing fault diagnosis under different loads. Because multi-scale capsule attention networks can improve feature expression ability and anti-noise performance, the fault data can be better expressed. Using the domain adaptation ability of joint distribution optimal transport, the feature distribution of fault data under different loads is aligned, and domain-invariant features are learned. Through experiments that investigate bearings fault diagnosis under different loads, the effectiveness of MSCAN-JDOT is verified; the fault diagnosis accuracy is higher than that of other methods. In addition, fault diagnosis experiment is carried out in different noise environments to demonstrate MSCAN-JDOT, which achieves a better anti-noise ability than other transfer learning methods.

Keywords: intelligent fault diagnosis; domain adaptation; multi-scale capsule attention network; joint distribution optimal transport; different working loads

check for updates

Citation: Sun, Z.; Yuan, X.; Fu, X.; Zhou, F.; Zhang, C. Multi-Scale Capsule Attention Network and Joint Distributed Optimal Transport for Bearing Fault Diagnosis under Different Working Loads. *Sensors* **2021**, *21*, 6696. https://doi.org/10.3390/s21196696

Academic Editor: Jongmyon Kim

Received: 21 September 2021
Accepted: 6 October 2021
Published: 8 October 2021

1. Introduction

Industrial mechanical systems are developing in the direction of complexity, precision and integration, and the tightness of mechanical equipment is increasing [1]. Therefore, operation state monitoring for mechanical equipment is becoming increasingly important. A bearing is the core component of rotating mechanical equipment, and its ability to operate is very important. Once the bearing fails, it will not only affect the normal operation of mechanical equipment but also cause serious or irreparable accidents and threaten the safety of personnel and property. As a popular fault diagnosis method, data-driven intelligent fault diagnosis has attracted a large number of researchers' attention in recent years [2]. Using a large volume of fault samples, data-driven fault diagnosis methods can learn the knowledge implicit in the data, and they are particularly effective for the complicated systems in which it is otherwise difficult to obtain accurate mathematical system models [3]. With the rapid development of intelligent mechanical systems, data can be collected at a high speed, which brings both industry and academia new opportunities and challenges [4]. Hence, it is meaningful to find an intelligent bearing fault diagnosis method with better performance.

Traditional machine learning fault diagnosis methods, for instance, decision tree, random forest [5] and support vector machines (SVMs) [6], require a complex manual feature extraction and selection process, which has a significant effect on the diagnosis accuracy. In recent years, the application of deep neural networks [7] in the wide field of fault diagnosis has gradually increased. Through an end-to-end method, the deep network-based method can avoid the manual feature extraction and selection process, which is time-consuming and overly dependent on experience. As a typical deep learning algorithm, convolutional neural networks (CNNs) are widely adopted in the broad area of fault diagnosis. For example, in [8], Wen et al. proposed a novel CNN algorithm based on the well-known Lenet-5 model for bearing fault diagnosis. In this CNN algorithm, one-dimensional raw vibration signals are skillfully processed and converted into two-dimensional grayscale images by signal superposition, and then the obtained grayscale images are fed into a CNN, which is used for fault diagnosis. Han et al. [9] presented a hybrid fault diagnosis framework, which mainly contains two parts: one part is a spatial-temporal pattern network, which is focused on the task of spatial–temporal feature learning, and the other part is a CNN, which is devoted to conditional classification. Zhang et al. [10] presented a well-designed CNN model, in which the first layer has wide convolution kernels, and the one-dimensional raw vibration signals are fed into the proposed CNN model. Experimental results indicated that the well-designed CNN model have good anti-noise performance in fault diagnosis. In addition, many CNN-based fault diagnosis methods using two-dimensional time-frequency image representations have also been exploited, such as [11]. To fully exploit the advantages of the well-trained CNN model in feature learning, the core idea of these methods is to convert one-dimensional time domain original training samples into time-frequency images for training network.

Compared with the fault diagnosis approaches based on classical shallow learning models, deep network-based methods show superior performance. Assuming that the data acquisition processes of training and testing sets are conducted under the same working condition, i.e., the data distribution of training and testing set is consistent, the majority of deep network-based fault diagnosis methods are effective. However, this strict assumption is nearly impossible in practical applications. For rolling bearings, the data distribution will change when the working load or rotating speed changes. To address this issue, transfer learning is an attractive alternative, which bring us a new perspective. In [12] and [13], Yan et al. proposed new fault diagnosis methods based on transfer learning, which promotes the research and application of transfer learning in the area of fault diagnosis significantly. In addition, extensive fault diagnosis experiments were conducted and experimental results indicated that the presented methods achieved impressive and promising performance.

As a representative and widely adopted transductive transfer learning method, domain adaption (DA) technique could align features distribution between the target domain and source domain during the training procedure while maintaining a good classification result. For example, using maximum mean discrepancy (MMD) and multi-kernel model, An et al. [14] presented a fault diagnosis framework, which achieved a high diagnosis accuracy. In [15], a new partial adversarial DA fault diagnosis approach was presented based on stacked auto-encoder. Using MMD and domain adversarial training (DAT), Li et al. [16] presented a novel diagnosis scheme, which achieved enhanced feature representation ability, and the ensemble learning was adopted to obtain the final diagnosis result. Wen et al. [17] proposed a new deep transfer model-based diagnosis approach, in which, the feature extraction task is fulfilled by a sparse auto-encoder network, and the inconsistency between the distributions of testing and training set is minimized by the MMD, thereby the domain adaptation process is accomplished. Li et al. [18] presented an end-to-end scheme that combines bidirectional signals and capsule networks to input horizontal and vertical vibration signals into the neural network. Using the proposed scheme, domain-invariant features can be learned from training samples collected under variable working conditions. To minimum the distribution differences across domains, Chen et al. [19] presented a well-designed transfer network-based multi-domain diagnosis scheme, which integrates a task-specific encoder network and DAT. In [20], based on the multi-scale multi-domain feature, an improved diagnosis scheme was designed, which is effective in dealing with the fault diagnosis problems under polytrophic working conditions. In order to extract domain-invariant features from the raw signals, Wang et al. [21] presented a deep adversarial domain adaptation network (DADAN), which uses DAT and the Wasserstein distance. By embedding the useful discriminative knowledge in the label predictions into the domain classifier, Yu et al. [22] proposed a powerful conditional DADAN, which can align the features distribution between the target and source domains better. In addition, a new loss function is introduced to better extract invariant and discriminative features. Li et al. [23] proposed a novel multi-layer domain adversarial graph convolutional network (DAGCN), which uses graph convolution to extract features, and uses domain adversarial training and maximum mean discrepancy to minimize distribution differences of target and source domain features. Huang et al. [24] presented a new promising deep adversarial capsule network (DACN), which can not only separate the composite fault into several single faults intelligently, but also generalize faults under certain working conditions into faults under other new working conditions. Using optimal transport (OT), Liu et al. [25] presented a novel diagnosis approach based on deep DA model. First, an improved auto-encoder network was used to learn class discrimination features. Second, domain-invariant features are extracted by minimizing OT cost function between target and source domains. Finally, the offline trained classifier is tested with the target domain samples. The results indicated that these methods achieve better fault diagnosis accuracy and domain adaptability. Compared with previous deep learning methods without domain adaptation, the methods with domain adaptation can maintain a good fault diagnosis effect when the working conditions change. Thus, domain adaptation is an effective method for fault diagnosis under different working conditions. However, the above methods usually consider fault diagnosis in the case where only small working condition changes occur without considering the influence of noise on fault diagnosis.

In this article, to further facilitate the accuracy of bearing fault diagnosis under different loads as well as the anti-noise ability of the fault diagnosis model, a transfer learning-based method using multi-scale capsule attention network and joint distribution optimal transport (MSCAN-JDOT) is proposed. The main contributions are summarized as follows:

- A new transfer learning-based fault diagnosis approach called MSCAN-JDOT is proposed which accepts raw vibration signal as input and can effectively perform end-to-end fault diagnosis without the time-consuming and experience-dependent manual feature extraction.
- The proposed MSCAN-JDOT adopts multi-scale capsule attention networks as feature extraction networks, which can better extract fault features, and uses joint distribution optimal transport for domain adaptation, which can effectively align the fault features under different loads.
- MSCAN-JDOT achieves high accuracy and strong anti-noise performance for bearing fault diagnosis under different working loads.

The rest of this article is organized as follows. Section 2 briefly introduces the theory of capsule networks and optimal transport. Section 3 describes the proposed MSCAN-JDOT in detail. Section 4 evaluates the performance of MSCAN-JDOT on the rolling bearing dataset. Section 5 concludes this article.

2. Capsule Network and Optimal Transport

2.1. Capsule Network

In the traditional CNN [26], features are transferred to the next layer through a pooling operation. The regional maximum or regional average is selected through max pooling or mean pooling. However, the spatial information will be lost inevitably in the process of the pooling operation. To overcome this problem, Sabour et al. [27] presented the capsule network. Capsule networks use capsules instead of neurons in the traditional neural network to extract invariant features more effectively. The capsule is a vector consisted of a certain number of neurons, and each neuron indicates a certain attribute of a particular instance, such as angle, color, and other properties [27]. After that, several improved capsule networks have been proposed. For example, Hinton et al. [28] proposed a matrix capsule network with EM routing algorithm. Ribeiro et al. [29] proposed a capsule routing algorithm based on Variational Bayes, which improved the routing mechanism of capsule network. Similar to other neural networks, the capsule network is also composed of multi-layer networks, primarily including the convolutional layer, primary capsule layer and digit capsule layer. Taking MNIST handwritten digit classification as an example, a simple three-layer capsule network is illustrated in Figure 1. The first convolution layer extracts the input image into a feature map, which serves as the input to the primary capsule layer. Second, the primary capsule layer extracts the low-level features and divides them into capsules with dimensions of 8. Then, the digit capsule layer obtains the output capsules from the primary capsule layer through dynamic routing and converts them into capsules with dimensions of 16. Finally, the capsules are classified using other layers and classifiers.

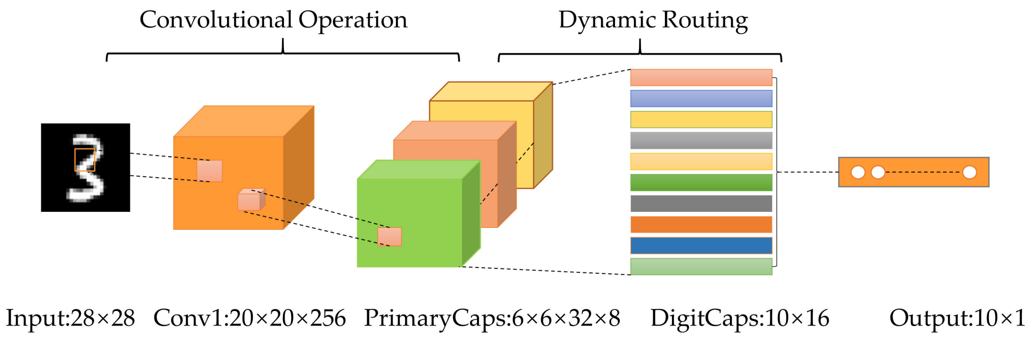


Figure 1. Architecture of a capsule network with three layers.

In capsule network, the length of the output capsule vector represents the probability of the existence for the instance category. Therefore, when the output category is consistent with the label, the output capsule has a long instantiation vector. Each digital capsule has a separate margin loss:

$$L_c = T_c \max\left(0, m^+ - \|\mathbf{v}_c\|\right)^2 + \lambda(1 - T_c)\max\left(0, \|\mathbf{v}_c\| - m^-\right)^2 \tag{1}$$

where c is the classification category, T_c is the indicator function of the classification, $T_c = 1$ when category c exists, $T_c = 0$ when category c does not exist. λ is the trade-off coefficient, m^+ is the upper boundary of classification probability and m^- is the lower boundary of classification probability. In addition, $\|\mathbf{v}_c\|$ is the L_2 distance of the vector $\mathbf{v}_c$. Since each category of capsule has a separate margin loss L_c, the total margin loss is the sum of L_c for all categories.

2.2. Optimal Transport

Optimal transport [30] (OT) is a method that can be used to compare probability distributions in a geometrically reasonable way. OT studies the empirical distribution and makes use of the geometric structure of the data embedding space. According to Equation (2), OT searches for a probabilistic coupling $\gamma \in \Pi(\mu_1, \mu_2)$ between two distributions μ_1 and μ_2, which finds a minimal transport cost:

$$OT_c(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1,\mu_2)} \int_{R^2} c(x_i, x_j) d\gamma(x_i, x_j) \tag{2}$$

In the discrete case, this becomes:

$$OT_c(\mu_1, \mu_2) = \min_{\gamma \in \Pi(\mu_1,\mu_2)} < \gamma, \mathbf{C} >_F \tag{3}$$

where x_i, x_j belong to μ_1, μ_2, respectively, the cost function $c(x_i, x_j)$ measures the difference between x_i and x_j, $\mathbf{C}$ is the cost matrix composed of $c(x_i, x_j)$, and $\Pi(\mu_1, \mu_2)$ describes the joint probability distribution of μ_1, μ_2.

Optimal transport has been used as a common method that situates the source distribution and target distribution closer to each other by finding a transmission probability coupling matrix γ between two different distributions to minimize the cost matrix $\mathbf{C}$. Moreover, experiments demonstrate that better constraint of the structure of γ using entropy or regularization terms contributes to better empirical results [31].

3. Proposed Method

In this article, to improve fault diagnosis accuracy under different loads as well as improve the anti-noise performance of the model, an intelligent fault diagnosis approach based on multi-scale capsule attention network and joint distribution optimal transport is proposed. Figure 2 shows the architecture of MSCAN-JDOT, which primarily contains four components: data input, feature extraction, classifier and domain adaptation. First, in the data input component, the model accepts a one-dimensional original sample as data input without any manual feature extraction. The labeled samples under one load are used as the source domain and the unlabeled samples under other loads are used as the target domain. The second component, feature extraction, includes a multi-scale convolution layer, an attention module, a primary capsule layer, a digit capsule layer and a fully connected layer. Then, in the classifier component, the fully connected layer transforms the feature dimensions and the fault diagnosis results are obtained using softmax. Finally, in the domain adaptive component, the adaptability of the model under different loads is implemented using joint distribution optimal transport.

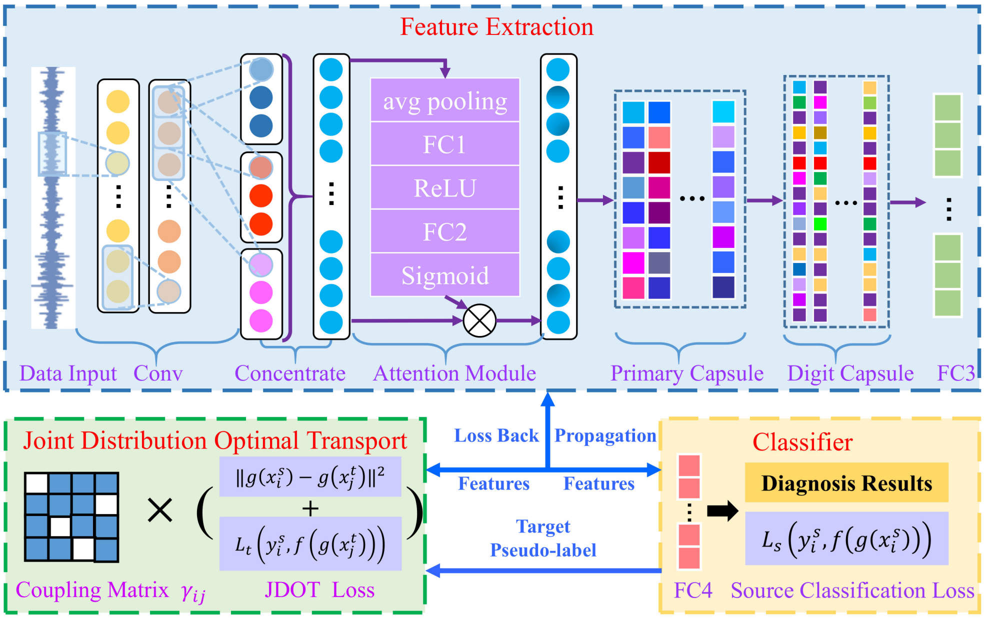


Figure 2. Architecture of the MSCAN-JDOT.

3.1. Feature Extraction Details

In feature extraction, a multi-scale capsule attention network is proposed. First, three one-dimensional convolution layers are applied to directly learn the feature representation from one-dimensional raw signals. The first convolution layer uses wide convolution kernels. Wide convolution, on the one hand, can expand the receptive field of convolution operation and accelerate the speed of model training; on the other hand, wide convolution can enhance the anti-noise ability of the model. The second convolution layer uses small size kernels to enhance the local feature extraction ability. The convolution process of the first two layers is described as follows:

$$y_k = \sigma(w_k \otimes x + b_k) \tag{4}$$

where y_k is the output of the kth layer, w_k and b_k represent the weights and bias of the convolutional process, x represents the input of the convolution layer, $\otimes$ indicates the convolutional calculation, and σ is the ReLU activation function. To better extract the domain-invariant features of fault data, multi-scale convolutional layer, which can be described as Equation (5), is added in the third layer of the model:

$$y_{ms} = concentrate(y_{31}, y_{32}, y_{33}) \tag{5}$$

where y_{31}, y_{32}, y_{33} are convolution outputs with convolution kernels of 3, 8 and 16, respectively, y_{ms} is the output of multi-scale convolution layer, and $concentrate(\cdot)$ indicates splicing by channel.

Second, a channel attention module, which can focus on more meaningful input channels, is added after the multi-scale convolution layer. To calculate channel attention effectively, average pooling is adopted to compress the spatial dimension of input features. Then, the full connection layer and sigmoid is used to calculate the attention weight on the channel. Finally, the attention weight is multiplied by the corresponding channel to obtain the input features with attention. The process can be described as follows:

$$y_a = x_a \cdot Sigmoid(W_{a_1} \cdot (\sigma(W_{a_2} \cdot pool(x_a) + b_{a_2})) + b_{a_1}) \tag{6}$$

where x_a is the input of the attention module, y_a is the output of the attention module, W_* and b_* are the weight and deviation of the full connection layer.

Third, a primary capsule layer and a digit capsule layer are added after the attention module, because the capsule network can extract various attributes of samples and better express the data features. The primary capsule layer can be described as follows:

$$y_{conv} = \sigma\left(W_{conv} \otimes x_{caps} + b_{conv}\right) \tag{7}$$

$$y_{pcaps} = \sigma\left(W_{pcaps} \otimes y_{conv} + b_{pcaps}\right) \tag{8}$$

where x_{caps}, y_{conv} are the corresponding input of the primary capsule layer, y_{pcaps} is the output of the primary capsule layer, W_* and b_* are the weights and bias in the corresponding layer.

Aiming to build the relationships between two capsule layers, dynamic routing algorithm, which is displayed in Algorithm 1, is introduced between the primary and digit capsule layers. In Algorithm 1, b_{ij} is the bias coefficient from capsule i in lth layer to capsule j in the next layer, and it is initialized to zero before algorithm iteration. $u_{j|i}$ is the intermediate prediction vector between the ith capsule and the jth capsule, and it is equal to the multiplication of u_i and the weight coefficient matrix W_{ij}. c_{ij} is the weight coefficient of the intermediate prediction vector $u_{j|i}$. b_i and c_i is the set of b_{ij} and c_{ij}. The squash function is similar to the activation function in the convolutional neural network that carries out nonlinear transformation on the input vector and compacts the input vector to [0, 1]:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \tag{9}$$

where s_j is the input to the squash function and v_j is the output of the squash function.

Algorithm 1 Dynamic routing algorithm

```
Procedure routing $(u_{j|i}, r, l)$
    for all capsule $i$ in layer $l$ and capsule $j$ in layer $l+1$: $b_{ij} \leftarrow 0$.
    for $r$ iterations do
        for all capsule $i$ in layer $l$: $c_i \leftarrow softmax(b_i)$
        for all capsule $j$ in layer $l+1$: $s_j \leftarrow \sum_i c_{ij} u_{j|i}$
        for all capsule $j$ in layer $l+1$: $v_j \leftarrow squash(s_j)$
        for all capsule $i$ in layer $l$ and capsule $j$ in layer $l+1$: $b_{ij} \leftarrow b_{ij} + u_{j|i} \cdot v_j$
    return $v_j$
```

The dynamic routing adopts an iteration number of 3, which was shown to be effective in [27]. In the first iteration, because b_{ij} is set to 0, all intermediate prediction vectors $u_{j|i}$ share the same weight coefficient c_{ij}. As the iteration proceeds, the intermediate prediction vector $u_{j|i}$, which is more similar to the high-level capsule v_j, has a larger weight coefficient c_{ij}. This coefficient ultimately ensures that the features of the low-level capsule are more likely to be transferred to a similar, high-level capsule. After dynamic routing, the output y^j_{dcaps} of digit capsule layer can be described as follows:

$$y^j_{dcaps} = v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|} \tag{10}$$

Finally, the result of feature extraction is obtained through a full connection layer. The model parameters of MSCAN-JDOT are shown in Table 1.

Table 1. MSCAN-JDOT model parameters.

Layer Name	Kernel Size	Filters	Strides	Padding	Capsule Dimension	Capsules Number	Output Shape
Input	-	-	-	-	-	-	(2048,2)
Conv1	64	16	1	same	-	-	(2048,16)
Conv2	32	32	8	valid	-	-	(253,32)
Conv3(multi-scale conv)	3/8/16	16/16/16	3	same	-	-	(253,48)
Attention avg_pooling	-	-	-	-	-	-	(253,48) (1,48)
FC1							(1,16)
FC2							(1,48)
Primary Capsule	3	256	1	valid	8	32	(32,8)
Digit Capsule	-	-	-		16	10	(10,16)
Flatten	-	-	-		-	-	(160)
FC3	-	-	-		-	-	(128)
FC4	-	-	-		-	-	(10)

3.2. *JDOT Domain Adaptation*

Courty et al. [31] presented the well-known joint distributed optimal transport (JDOT) method to avoid two-step adaptation (i.e., first performing domain feature adaptation and then learning the classifier from the adaptive features) by diametrically learning the classifier embedded in the cost function c. The goal of JDOT is to align the joint distribution of data features and labels, rather than just aligning the feature distribution. In the domain adaptive task, it is assumed that (x^s_i, y^s_i) and (x^t_j, y^t_j) are samples from the source and target domain, respectively. The JDOT cost function consists of the following two parts:

$$d\left(x^s_i, y^s_i; x^t_j, y^t_j\right) = \alpha c\left(x^s_i, x^t_j\right) + \lambda_t L\left(y^s_i, y^t_j\right) \tag{11}$$

where $c\left(x^s_i, x^t_j\right)$ is the cost function that aligns the feature distribution, $L\left(y^s_i, y^t_j\right)$ is the cost function that aligns the label distribution, and α and λ_t are two coefficients that weight the cost of the two parts. Usually, the target domain label y^t_j is unknown and the pseudo-label

$f\left(g\left(x_j^t\right)\right)$ is generated using the classifier f and feature extractor g. Therefore, the JDOT objective function becomes:

$$\inf_{f,\gamma\in\Pi(\mu_1,\mu_2)}\int_{R^2}\alpha c\left(x_i^s,x_j^t\right)+\lambda_t L\left(y_i^s,y_j^t\right)d\gamma(x_1,x_2) \tag{12}$$

In the discrete case, the objective function becomes:

$$\min_{f,\gamma\in\Pi(\mu_s,\mu_t)} < \gamma,\mathbf{D}_f >_F \tag{13}$$

where $\mathbf{D}_f$ is the set of $d\left(x_i^s,y_i^s;x_j^t,y_j^t\right)$.

The proposed MSCAN-JDOT method aligns the features distribution of source domain and target domain using joint distribution optimal transport; additionally, the label distribution is considered while the feature distribution is aligned. The objective function of MSCAN-JDOT can be described as:

$$\min_{\gamma,f,g}\frac{1}{n^s}\sum_i L_s(y_i^s,f(g(x_i^s)))+\sum_{i,j}\gamma_{ij}\left(\alpha\left\|g(x_i^s)-g\left(x_j^t\right)\right\|^2+\lambda_t L_t\left(y_i^s,f\left(g\left(x_j^t\right)\right)\right)\right) \tag{14}$$

where $L_s\left(y_i^s,f\left(g\left(x_i^s\right)\right)\right)$ is the source classification loss, $\left\|g\left(x_i^s\right)-g\left(x_j^t\right)\right\|^2$ and $L_t\left(y_i^s,f\left(g\left(x_j^t\right)\right)\right)$ is feature alignment loss and label alignment loss between the source domain and target domain, γ is the coupling matrix, f is the classifier, g is the feature extractor, and α, λ_t are two coefficients weighting the loss of the two parts. The parameters α and λ_t are set as $\alpha=0.001$ and $\lambda_t=0.0001$ according to [32].

3.3. General Procedure of the Proposed Method

The flowchart is shown in Figure 3 and the general procedures of the proposed MSCAN-JDOT are summarized as follows:

1. Data Input: In this step, the raw data sampled under different working loads are split into target domain and source domain. The training sets contains the labeled source domain samples and the unlabeled target domain samples, while the testing sets only contains the unlabeled target samples.
2. Training Stage: In this step, the training samples are input to the feature extraction network, and then the domain adaptation aligns the features of the source domain and target domain. Through the source prediction labels and target pseudo-labels generated by the classifier, the whole loss function of MSCAN-JDOT can be calculated by Equation (13). Finally, the model parameters can be updated with backward propagation.
3. Testing Stage: In this step, testing samples are used to validate the performance of the MSCAN-JDOT, which is well trained after sufficient epochs. In this stage, the network only carries out forward propagation without backward propagation. The model is evaluated by label prediction results and features alignment effect.

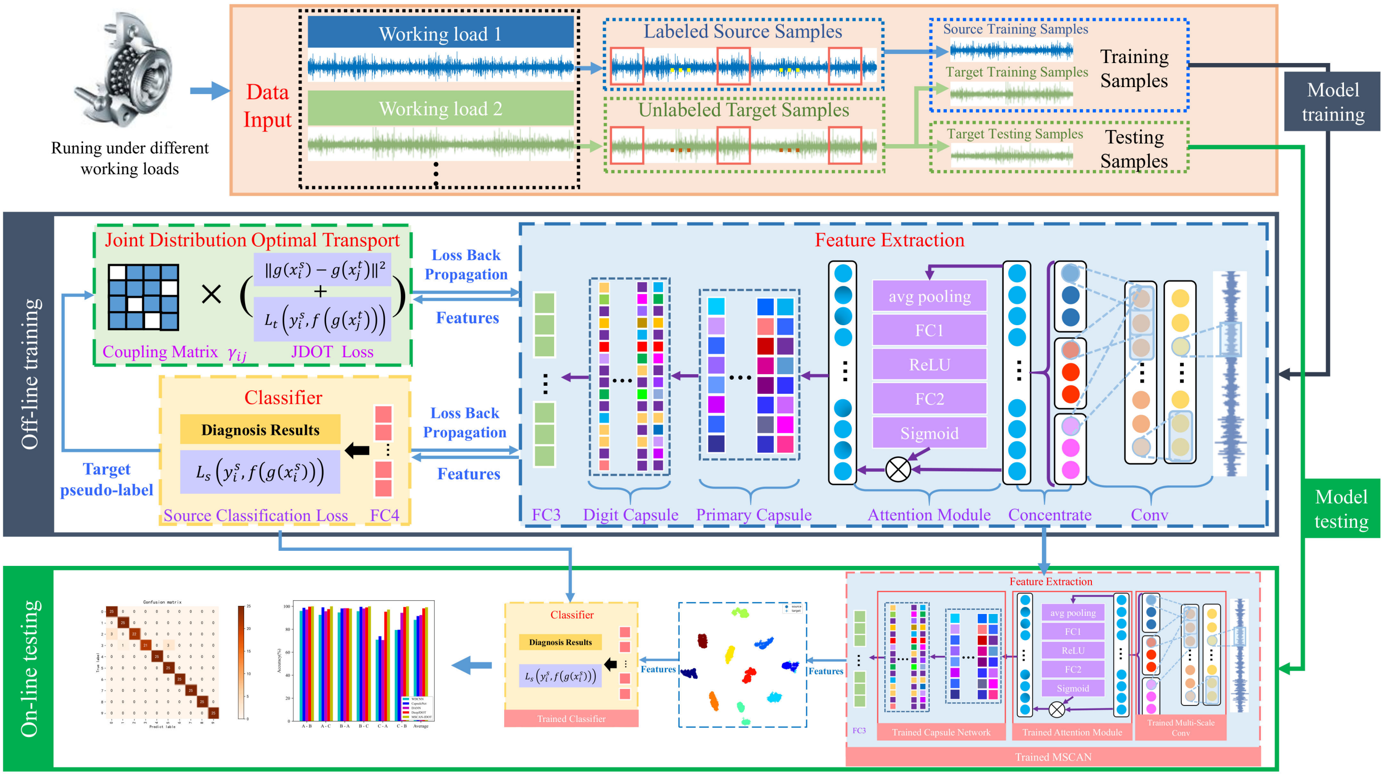


Figure 3. Flowchart of the MSCAN-JDOT.

4. Experimental Analysis

In the actual operation of rolling bearings, the load and bearing speed will inevitably change. This section evaluates the proposed MSCAN-JDOT model under different loads. Since noise is inevitable in actual working environment and vibration signals are easily disturbed by noise, the anti-noise performance of the MSCAN-JDOT model is also evaluated.

4.1. Dataset Introduction and Dataset Split

4.1.1. Dataset Introduction

This article uses the Case Western Reserve University (CWRU) dataset [33] as the experimental data. An accelerometer is used for data collection. The sampling frequency is 12 kHz, and the sampling time is 10 s. Each dataset has 120,000 sampling points. This experiment uses three different load datasets, as shown in Table 2. A, B, and C represent 1, 2, and 3 loads, respectively, and the speed gradually decreases as the load increases. For each load, there are three different fault diameters and three different fault locations, for a total of nine different fault types. Each fault dataset contains three sets of data from different sampling locations, including driver-end data, fan-end data, and basic data. Table 3 shows the label assignment of the nine fault data and normal data. Figure 4 shows the drive-end data of 0.014_Ball under three different loads, and the data distribution changes significantly under different loads.

Table 2. Different load datasets of CWRU.

Dataset Name	Speed (rpm)	Load (HP)	Fault Diameter	Fault Location
A	1772	1	0.007,0.014,0.021	Ball, InnerRace, OuterRace
B	1750	2	0.007,0.014,0.021	Ball, InnerRace, OuterRace
C	1730	3	0.007,0.014,0.021	Ball, InnerRace, OuterRace

Table 3. Label assignment.

Health Conditions	Label
Normal	0
0.007_Ball	1
0.007_InnerRace	2
0.007_OuterRace	3
0.014_Ball	4
0.014_InnerRace	5
0.014_OuterRace	6
0.021_Ball	7
0.021_InnerRace	8
0.021_OuterRace	9

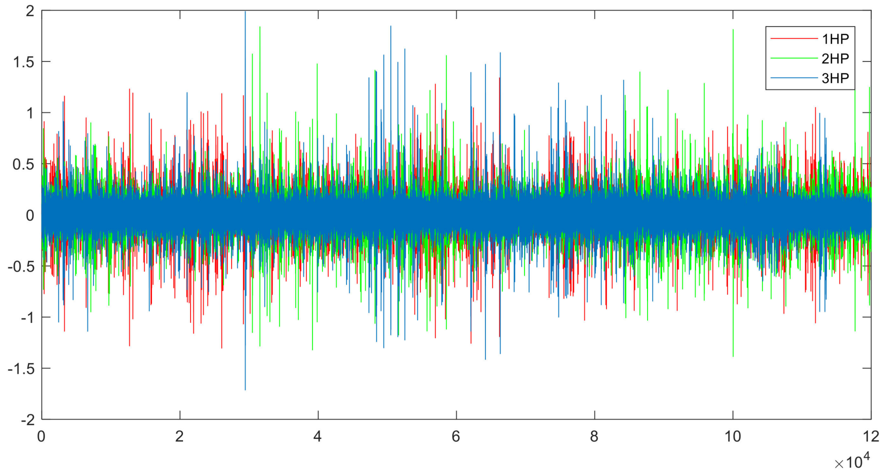


Figure 4. Drive-end data of 0.014_Ball under three different loads.

4.1.2. Dataset Split

Each original dataset contains 120,000 sampling points, which are separated into two groups with the same size. Because the number of samples in the training set is small, overlapping sampling is adopted for data splitting. The length and stride of sliding window affect the representation ability of fault attributes, so it is important to select the appropriate length and stride. As can be seen from Figure 5, the sliding stride has a great influence on the effectiveness of the method. With the increase in stride, the accuracy increases at first and then decreases. When the sliding stride is larger than 80, the performance of the method decreases sharply. Because the total number of training samples is small when the stride is too large, the effect of overlapping sampling is poor. The effect of sliding window length on the performance of this method is relatively small. When the length of sliding window is 2048, the overall performance is better. Therefore, the sliding window length is set to 2048 and the sliding stride is set to 80. Overlapping sampling is not adopted for the test set, and the length of each data sample is also 2048 sampling points. The data split is shown in Figure 6. Finally, 660 training samples and 25 test samples generated during each fault data collection step are selected. There are a total of 19,800 training samples and 750 test samples. In this article, each input data sample selects driver-end and fan-end data. Therefore, the input data dimension is (2048, 2). Before neural network training, input data need to be normalized:

$$x^* = \frac{x - \mu}{\sigma} \tag{15}$$

where μ and σ is the sample mean and deviation, respectively.

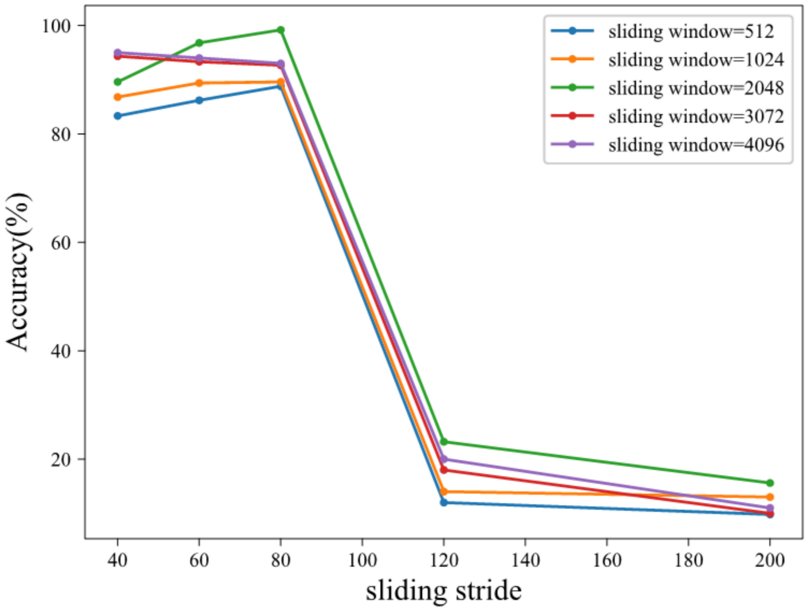


Figure 5. Sliding window length and stride parameter selection.

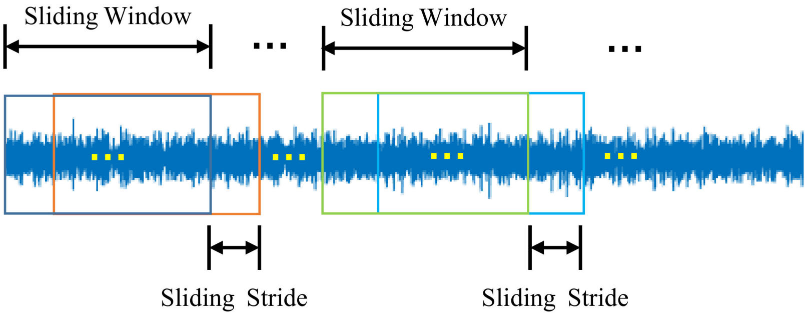


Figure 6. Data split.

4.2. *Experimental Results and Performance Analysis*

4.2.1. Fault Diagnosis Experiments under Different Loads

In this part of the experiment, the source domain data with labels and the target domain data without labels are used as input, and the trained model is evaluated on the test set from the target domain without labels. According to the datasets and data split methods of three different loads, six experiments are used to test the MSCAN-JDOT model, including A-B, A-C, B-A, B-C, C-A, and C-B. For example, A-B represents that these experimental models are trained using dataset A with labels and dataset B without labels and tested with dataset B.

To prove the effectiveness of the proposed MSCAN-JDOT, three widely applied algorithms are selected as competitors: the WDCNN [10], CapsuleNet, DANN [34], and DeepJDOT [32]. The WDCNN is the first layer wide convolutional kernel convolutional neural network. The CapsuleNet uses a one-dimensional capsule network for feature extraction, and the parameters are the same as those of capsule network in MSCAN-JDOT. The DANN is a domain adversarial neural network in which a 1D-CNN is used to extract features and an adversarial network is used to self-adaptively align the feature distribution. DeepJDOT uses a 1D-CNN for feature extraction and JDOT for domain adaptation. In WDCNN, DANN and DeepJDOT, the feature extraction component has the same structure, but the domain adaptation is different. In the feature extraction component, DeepJDOT is different from MSCAN-JDOT, but the domain adaptive method is the same. All experimental input data are the same, and the model trained over 100 epochs. Table 4 shows the transfer accuracy under different loads. Figure 7 is a visualization of Table 4, which more intuitively shows the effectiveness of the MSCAN-JDOT model.

Table 4. Transfer accuracy on CWRU dataset (%).

CWRU	A-B	A-C	B-A	B-C	C-A	C-B	AVG
WDCNN	96.04	92.60	94.68	95.76	70.96	79.44	88.25
CapsuleNet	98.76	99.16	98.16	99.60	73.96	79.60	91.54
DANN	97.20	95.72	98.40	98.12	70.64	94.36	92.41
DeepJDOT	99.72	97.60	98.32	99.08	95.32	99.44	98.25
MSCAN-JDOT	100	100	98.00	100	97.20	100	99.20

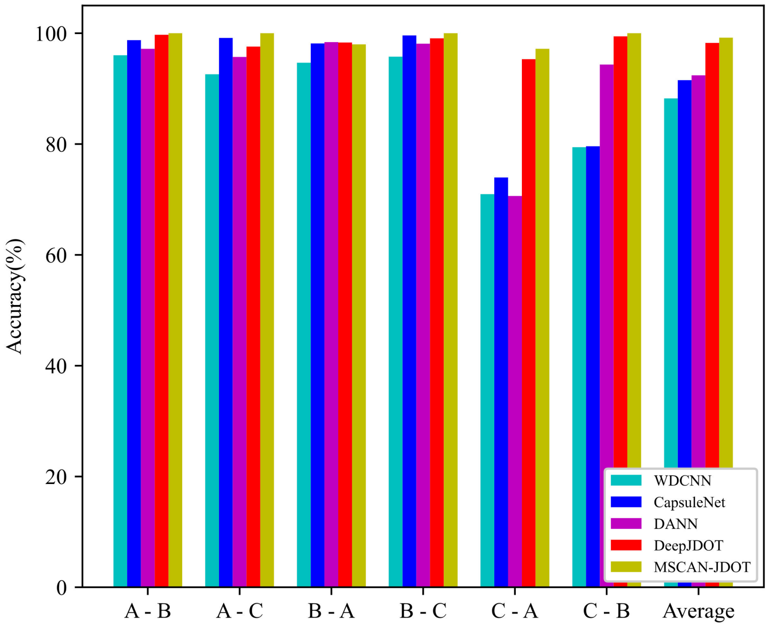
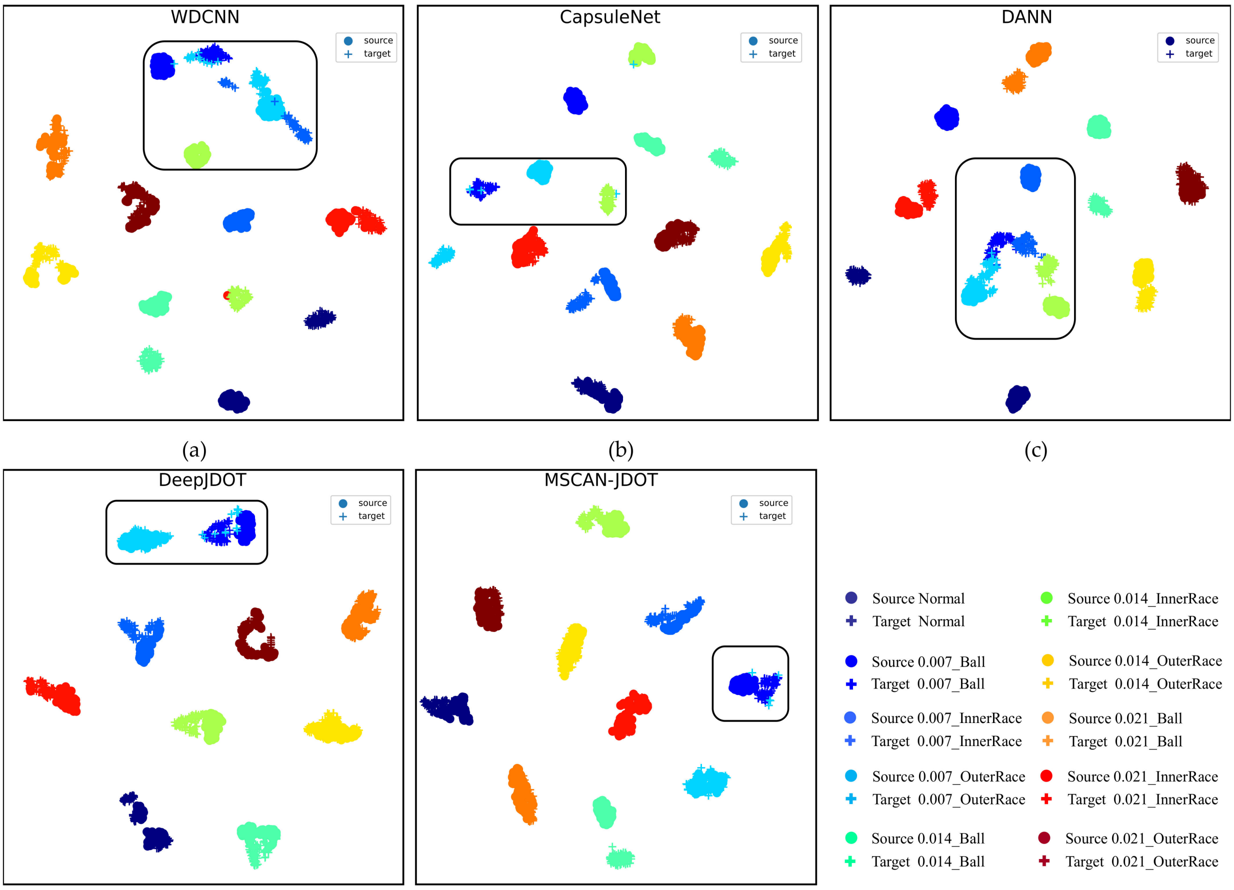


Figure 7. Visualization of transfer results on CWRU dataset.

4.2.2. Analysis of Fault Diagnosis Experimental Results under Different Loads

As seen from the experimental results in Section 4.2.1, the WDCNN has the worst diagnosis performance, and its average accuracy is 88.25%, and the accuracies for C-A and C-B are 70.96% and 79.44%, respectively. The average accuracy of the CapsuleNet is 91.54%, which is better than WDCNN. The DANN adopts an adversarial network for domain adaptation, and the fault diagnosis performance under different loads are improved to a certain extent. The average accuracy is 92.41%, while the accuracy is only 70.64% for C-A, which is similar to the WDCNN. However, the accuracy is the highest among the four models for B-A. MSCAN-JDOT and DeepJDOT adopt joint distribution optimal transport and demonstrate a significant improvement when compared with other domain adaptive methods. The average fault diagnosis accuracy is approximately 10% higher compared with WDCNN and 6% higher than that of the CapsuleNet and DANN. DeepJDOT uses convolutional networks and joint distribution optimal transport to achieve an average accuracy of 98.25% under different loads and 95.32% for C-A. The proposed MSCAN-JDOT uses a multi-scale capsule attention network and joint distribution optimal transport; this combination improves the fault diagnosis performance most obviously. The average accuracy of MSCAN-JDOT reaches 99.20%, which is the highest among the four models. In addition, the accuracy of MSCAN-JDOT is the highest for A-B, A-C, B-C, C-A, and C-B. It can be proved from the above experimental results that the feature extraction effect of the multi-scale capsule attention network is better than that of the convolutional neural network. The input data use two sets of data for each fault, which equivalently increases the number of fault attributes at different sampling locations. The output of the multi-scale capsule attention network is a vector, which can better extract complex fault features and other fault attributes.

To better verify the performance of MSCAN-JDOT, the features extracted by feature extraction, that is, the domain-invariant features, are reduced to two dimensions using t-sne and visualized. The result for C-A is illustrated in Figure 8, where the source domain data are represented by "·", the target domain data are represented by "+", and ten different colors represent ten types of faults. From the data in the boxes in Figure 8a–c, it can be seen that the domain adaptation and fault classification effects of the WDCNN, CapsuleNet and DANN are relatively poor. On the one hand, the distance of features between the target domain and source domain data is sizeable; on the other hand, there is a large amount of overlap between different color blocks. As seen from the data in the box in Figure 8d, DeepJDOT shows some improvement when compared with the WDCNN, CapsuleNet and DANN. The distance of features between the target domain and source domain data is relatively small, and there is less overlap between different color blocks. Figure 8e indicates that the feature alignment effect of the proposed MSCAN-JDOT is similar to that of DeepJDOT, but MSCAN-JDOT has the best fault classification effect, and there is almost no overlap between different color blocks.

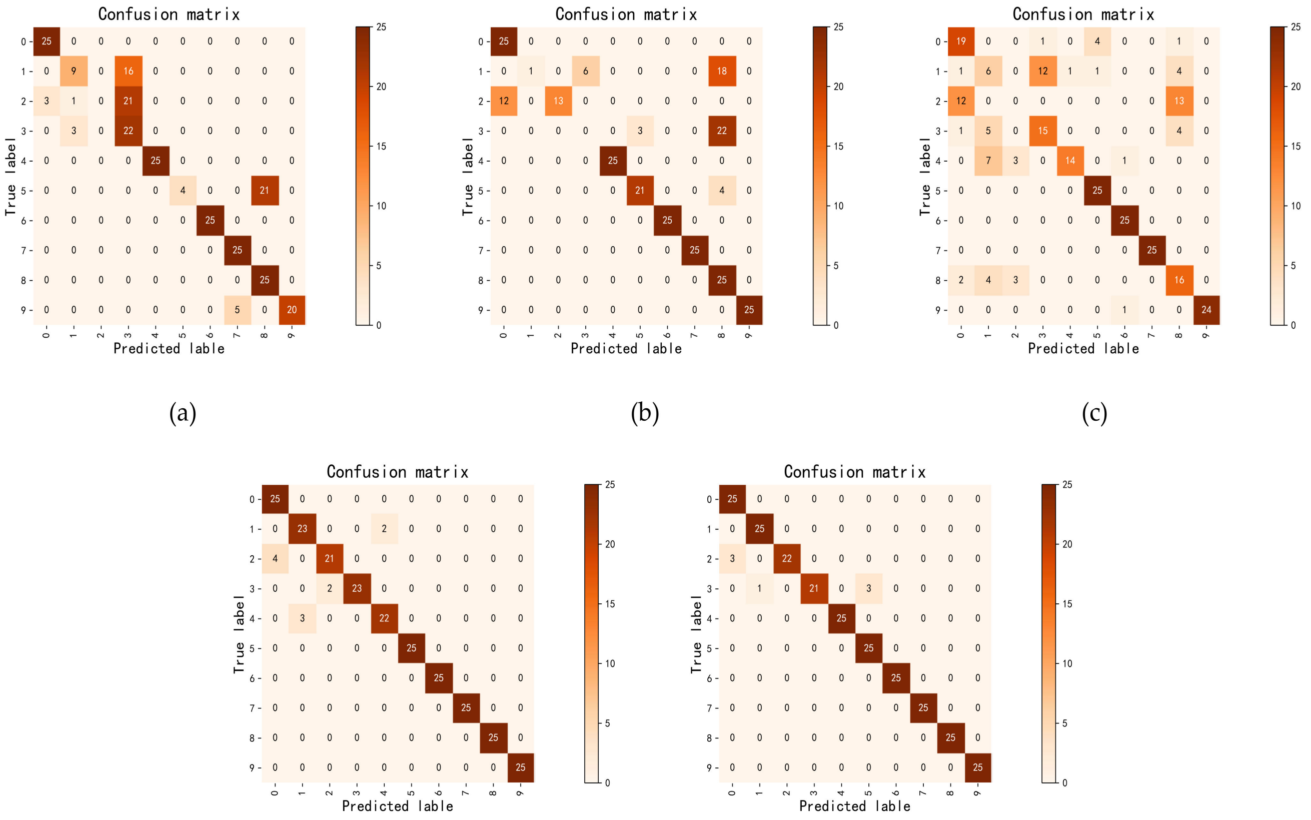


(a) (b) (c)

(d) (e)

Figure 8. Visualization results of C-A domain-invariant features. Domain-invariant features of (**a**) WDCNN, (**b**) CapsuleNet, (**c**) DANN, (**d**) DeepJDOT, and (**e**) MSCAN-JDOT.

Figure 9 illustrates the confusion matrices obtained by different models using the C-A test set, where the horizontal axis is the predicted label, the vertical axis is the true label, and the diagonal elements are the quantity of samples correctly classified, and other positions are the misclassified samples. As seen from Figure 9a–c, the WDCNN, CapsuleNet and DANN have a large number of incorrectly predicted samples. As shown in Figure 9d, the number of incorrectly predicted samples from DeepJDOT is significantly reduced. The confusion matrix of MSCAN-JDOT is illustrated in Figure 9e, from which we can see that the number of incorrectly predicted samples is the smallest. This result indicates that MSCAN-JDOT has the best classification performance, which is consistent with the transfer accuracy in Figure 7 and the t-sne visualization results in Figure 8.



(a)

(b)

(c)

(d)

(e)

Figure 9. Confusion matrix from C-A test results. Confusion matrix of (**a**) WDCNN, (**b**) CapsuleNet, (**c**) DANN, (**d**) DeepJDOT, and (**e**) MSCAN-JDOT.

4.2.3. Anti-Noise Experiments under Different Levels of Noise

In this section, MSCAN-JDOT's anti-noise performance is evaluated. In the actual production environment, noise is inevitable. Hence, Gaussian white noise is added to the raw vibration signal to simulate actual noise. Thus, composite vibration signals are obtained with different signal–noise ratios (SNRs). The SNR is defined as follows:

$$\mathrm{SNR_{dB}} = 10\log_{10}\left(\frac{S}{N}\right) \tag{16}$$

where S is the power of the raw signal and N is the power of the noise signal. The larger the SNR, the smaller the noise signal. SNR = 0 means that the power of the noise signal is the same as that of the raw signal.

In this experiment, the anti-noise performance is analyzed using A-C. Noise is added to target domain C while no noise is added to source domain A, and other parameters are unchanged from those used in the fault diagnosis experiments under different loads, which are introduced in Section 4.2.1. The noise-free fault data and noisy data are shown in Figure 10. As shown in Figure 10, with the reduction in noise, the distribution of the noisy fault data becomes increasingly similar to the distribution of the noise-free fault data.

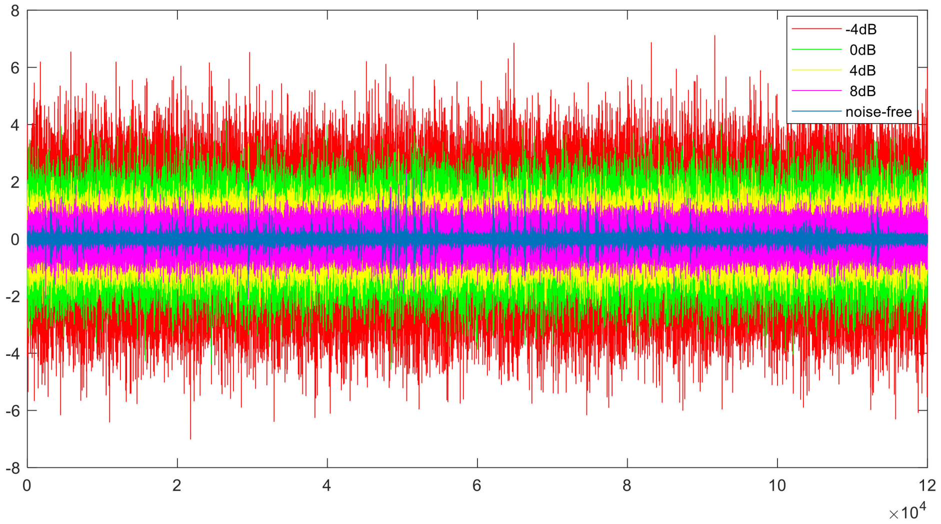


Figure 10. 0.014_ball noise-free and noisy data under 3HP.

Table 5 shows the transfer accuracy from A-C; the signal–noise ratio ranges from -4dB to 8dB. Figure 11 is a visualization of Table 5, and displays the anti-noise performance of MSCAN-JDOT.

Table 5. A-C transfer accuracy in noisy environment (%).

SNR(dB)	−4	−2	0	2	4	6	8
DANN	25.64	24.20	23.64	24.48	27.60	32.76	40.40
DeepJDOT	32.80	36.64	42.36	46.72	54.72	56.64	61.96
MSCAN-JDOT	34.80	38.44	43.60	51.08	55.20	57.20	63.72

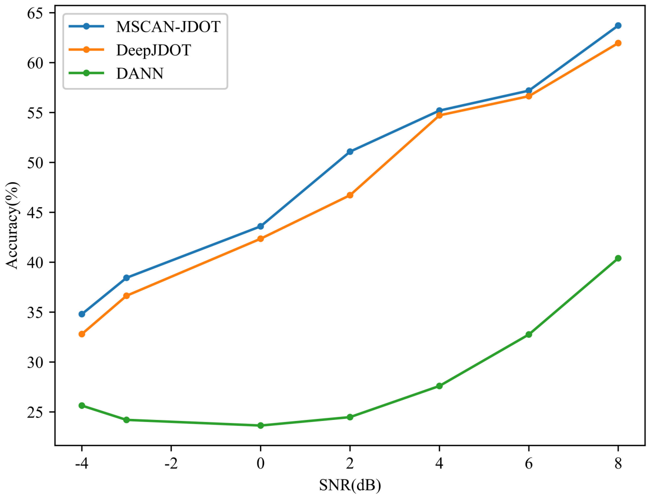


Figure 11. Comparison of A-C transfer results in noisy environment.

4.2.4. Anti-Noise Performance Analysis

According to the results of anti-noise experiment in Section 4.2.3, the proposed MSCAN-JDOT has a higher accuracy compared with other methods under different noise environments. The DANN performed the worst under noisy conditions, with a maximum accuracy of only approximately 40%. Compared with the DANN, DeepJDOT demonstrates a certain improvement in anti-noise performance, with a maximum accuracy of 61.96%. The proposed MSCAN-JDOT achieves 63.72% fault diagnosis accuracy when the SNR equals to 8 dB. Under -4 dB noise, the transfer accuracy of the DANN is less than 30%, the accuracy of DeepJDOT is 32.80%, and the proposed MSCAN-JDOT's accuracy is 34.80%, the highest accuracy among all tested methods. In the proposed method, the first layer adopts wide convolution kernel and the third layer adopts multi-scale operation. The wide convolution kernel and the multi-scale operation have certain anti-noise ability. In addition, the improved capsule network can better extract multiple attribute features of data by using capsules instead of neurons. Therefore, the features of noise-free data and noisy data can be aligned as much as possible through the domain adaptive module, so as to make the data features in noisy environment more distinguishable. It can also be seen from the results in Figure 11, the anti-noise performance of the proposed method is better than other methods. The above experimental results prove that the multi-scale capsule attention network and JDOT can improve the anti-noise performance. Although the anti-noise ability of the proposed MSCAN-JDOT is enhanced compared with that of other methods, the overall accuracy is not high. This is because MSCAN-JDOT does not add other anti-noise methods.

5. Conclusions

To improve fault diagnosis performance for rolling bearings under different loads, this article proposes a transfer learning fault diagnosis method based on multi-scale capsule attention network and joint distribution optimal transport. In this proposed method, the raw, one-dimensional vibration signal is used as input, and the fault features are extracted using the multi-scale capsule attention network. Joint distribution optimal transport is used for fault data domain adaptation under different loads. The proposed MSCAN-JDOT achieves outstanding performance in fault diagnosis under different working loads, with an average accuracy of 99.20%, which is better than that of other transfer learning methods. To address the impact of noise in the actual environment, the anti-noise performance of MSCAN-JDOT is also analyzed in this article. Under seven different noise conditions, the proposed method's fault diagnosis accuracy is also better than that of other transfer learning methods. The above experiments verify the excellent performance of the proposed MSCAN-JDOT.

In future work, the proposed fault diagnosis method will be further improved, and fault diagnosis under different loads in high-noise conditions will be studied to improve the generalization ability.

Author Contributions: Conceptualization, Z.S.; Formal analysis, Z.S. and X.Y.; Investigation, X.Y.; Methodology, Z.S., X.Y. and X.F.; Resources, Z.S. and X.F.; Software, Z.S.; Validation, Z.S. and X.Y.; Visualization, Z.S.; Writing—original draft, Z.S. and X.F.; Writing—review and editing, X.Y., F.Z. and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Natural Science Foundation of China (Grant No. 61803227, 61773242, 61973184), Key R & D Program of China (Grant No. 2017YFB1302400), Major Agricultural Applied Technological Innovation Projects of Shandong Province(SD2019NJ014), Independent Innovation Foundation of Shandong University (Grant No. 2018ZQXM005), Young Scholars Program of Shandong University, Weihai (Grant No. 20820211010).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to convey our deep appreciation to the editors and the reviewers for their insightful comments and constructive suggestions, which were very helpful in the improvement of our manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chen, Z.; Gryllias, K.; Li, W. Mechanical fault diagnosis using Convolutional Neural Networks and Extreme Learning Machine. *Mech. Syst. Signal Process.* **2019**, *133*, 106272. [CrossRef]
2. Shao, S.; Yan, R.; Lu, Y.; Wang, P.; Gao, R.X. DCNN-Based Multi-Signal Induction Motor Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 2658–2669. [CrossRef]
3. Xu, G.; Liu, M.; Jiang, Z.; Shen, W.; Huang, C. Online Fault Diagnosis Method Based on Transfer Convolutional Neural Networks. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 509–520. [CrossRef]
4. Ma, M.; Mao, Z. Deep-Convolution-Based LSTM Network for Remaining Useful Life Prediction. *IEEE Trans. Ind. Inform.* **2021**, *17*, 1658–1667. [CrossRef]
5. Wang, Z.; Zhang, Q.; Xiong, J.; Xiao, M.; Sun, G.; He, J. Fault Diagnosis of a Rolling Bearing Using Wavelet Packet De-noising and Random Forests. *IEEE Sens. J.* **2017**, *17*, 5581–5588. [CrossRef]
6. Shi, Q.; Zhang, H. Fault Diagnosis of an Autonomous Vehicle with an Improved SVM Algorithm Subject to Unbalanced Datasets. *IEEE Trans. Ind. Electron.* **2021**, *68*, 6248–6256. [CrossRef]
7. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A.K. Applications of machine learning to machine fault diagnosis: A re-view and roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [CrossRef]
8. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method. *IEEE Trans. Ind. Electron.* **2017**, *65*, 5990–5998. [CrossRef]
9. Han, T.; Liu, C.; Wu, L.; Sarkar, S.; Jiang, D. An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems. *Mech. Syst. Signal Process.* **2019**, *117*, 170–187. [CrossRef]
10. Zhang, W.; Peng, G.; Li, C.; Chen, Y.; Zhang, Z. A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals. *Sensors* **2017**, *17*, 425. [CrossRef] [PubMed]
11. Ma, P.; Zhang, H.; Fan, W.; Wang, C.; Wen, G.; Zhang, X. A novel bearing fault diagnosis method based on 2D image representation and transfer learning-convolutional neural network. *Meas. Sci. Technol.* **2019**, *30*, 055402. [CrossRef]
12. Shao, S.; McAleer, S.; Yan, R.; Baldi, P. Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning. *IEEE Trans. Ind. Inform.* **2019**, *15*, 2446–2455. [CrossRef]
13. Yan, R.; Shen, F.; Sun, C.; Chen, X. Knowledge Transfer for Rotary Machine Fault Diagnosis. *IEEE Sens. J.* **2020**, *20*, 8374–8393. [CrossRef]
14. An, Z.; Li, S.; Wang, J.; Xin, Y.; Xu, K. Generalization of deep neural network for bearing fault diagnosis under different working conditions using multiple kernel method. *Neurocomputing* **2019**, *352*, 42–53. [CrossRef]
15. Liu, Z.-H.; Lu, B.-L.; Wei, H.-L.; Chen, L.; Li, X.-H.; Wang, C.-T. A Stacked Auto-Encoder Based Partial Adversarial Domain Adaptation Model for Intelligent Fault Diagnosis of Rotating Machines. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6798–6809. [CrossRef]
16. Li, Y.; Song, Y.; Jia, L.; Gao, S.; Li, Q.; Qiu, M. Intelligent Fault Diagnosis by Fusing Domain Adversarial Training and Maximum Mean Discrepancy via Ensemble Learning. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2833–2841. [CrossRef]
17. Wen, L.; Gao, L.; Li, X. A New Deep Transfer Learning Based on Sparse Auto-Encoder for Fault Diagnosis. *IEEE Trans. Syst. Man, Cybern. Syst.* **2017**, *49*, 136–144. [CrossRef]
18. Li, L.; Zhang, M.; Wang, K. A fault diagnostic scheme based on capsule network for rolling bearing under different rotational speeds. *Sensors* **2020**, *20*, 1841. [CrossRef] [PubMed]
19. Chen, Z.; He, G.; Li, J.; Liao, Y.; Gryllias, K.; Li, W. Domain Adversarial Transfer Network for Cross-Domain Fault Diagnosis of Rotary Machinery. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 8702–8712. [CrossRef]
20. Lei, Z.; Wen, G.; Dong, S.; Huang, X.; Zhou, H.; Zhang, Z.; Chen, X. An Intelligent Fault Diagnosis Method Based on Domain Adaptation and Its Application for Bearings Under Polytropic Working Conditions. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–14. [CrossRef]
21. Wang, Y.; Sun, X.; Li, J.; Yang, Y. Intelligent Fault Diagnosis with Deep Adversarial Domain Adaptation. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–9. [CrossRef]
22. Yu, X.; Zhao, Z.; Zhang, X.; Sun, C.; Gong, B.; Yan, R.; Chen, X. Conditional Adversarial Domain Adaptation with Discrimination Embedding for Locomotive Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12. [CrossRef]
23. Li, T.; Zhao, Z.; Sun, C.; Yan, R.; Chen, X. Domain Adversarial Graph Convolutional Network for Fault Diagnosis Under Variable Working Conditions. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–10.
24. Huang, R.; Li, J.; Liao, Y.; Chen, J.; Wang, Z.; Li, W. Deep Adversarial Capsule Network for Compound Fault Diagnosis of Machinery Toward Multidomain Generalization Task. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–11. [CrossRef]
25. Liu, Z.; Jiang, L.; Wei, H.; Chen, L.; Li, X. Optimal Transport Based Deep Domain Adaptation Approach for Fault Diagnosis of Rotating Machine. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12.
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process.* **2012**, *2*, 1097–1105. [CrossRef]

27. Sabour, S.; Frosst, N.; Hinton, G. Dynamic routing between capsules. In Proceedings of the 31st Conference on Neural Information Processing System, Long Beach, CA, USA, 4–9 December 2017; pp. 3857–3867.
28. Hinton, G.; Sabour, S.; Frosst, N. Matrix capsules with EM routing. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
29. Ribeiro, F.D.S.; Leontidis, G.; Kollias, S. Capsule Routing via Variational Bayes. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 3749–3756.
30. Courty, N.; Flamary, R.; Tuia, D.; Rakotomamonjy, A. Optimal Transport for Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1853–1865. [CrossRef] [PubMed]
31. Courty, N.; Flamary, R.; Habrard, A.; Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In Proceedings of the Conference on Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3731–3740.
32. Damodaran, B.B.; Kellenberger, B.; Flamary, R.; Tuia, D.; Courty, N. DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Springer International Publishing: Cham, Switzerland, 2018; Volume 11208, pp. 467–483.
33. Case Western Reserve University Bearing Data Center Website. Available online: https://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website (accessed on 29 March 2021).
34. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2017**, *17*, 1–35.