Multi-Layer Domain Adaptation Method for Rolling Bearing Fault Diagnosis

Xiang Li [a,1] Qian Ding [b] Jian-Qiao Sun [c]

[a] *College of Sciences, Northeastern University, Shenyang 110819, China*

[b] *Department of Mechanics, Tianjin University, Tianjin 300072, China*

[c] *School of Engineering, University of California, Merced, CA 95343, USA*

Abstract

In the past years, data-driven approaches such as deep learning have been widely applied on machinery signal processing to develop intelligent fault diagnosis systems. In real-world applications, domain shift problem usually occurs where the distribution of the labeled training data, denoted as source domain, is different from that of the unlabeled testing data, known as target domain. That results in serious diagnosis performance degradation. This paper proposes a novel domain adaptation method for rolling bearing fault diagnosis based on deep learning techniques. A deep convolutional neural network is used as the main architecture. The multi-kernel maximum mean discrepancies (MMD) between the two domains in multiple layers are minimized to adapt the learned representations from supervised learning in the source domain to be applied in the target domain. The domain-invariant features can be efficiently extracted in this way, and the cross-domain testing performance can be significantly improved. Experiments on a popular rolling bearing dataset are carried out to validate the effectiveness of the domain adaptation approach, and the diagnosis performance is extensively evaluated in different scenarios. Comparisons with other approaches and related works on the same dataset demonstrate the superiority of the proposed method. The experimental results of this study suggest the proposed domain adaptation method offers a new and promising tool for intelligent fault diagnosis.

Key words: Fault diagnosis, Domain adaptation, Deep learning, Maximum mean discrepancy, Rolling bearing, Convolution neural network.

[1] Corresponding author. xiangli@mail.neu.edu.cn

1 Introduction

Rolling element bearings are critical components in heavy-duty machineries, manufacturing systems *etc.* and have been widely applied in modern industries. Unexpected bearing faults during long-term operations lead to large costs of maintenance and loss of safety [1]. In the past decades, bearing fault diagnosis has received considerable attention from researchers, and a large number of fault diagnosis methods have been proposed [2–8]. Especially, data-driven intelligent fault diagnosis methods, which are able to rapidly and efficiently process collected signals, provide reliable fault diagnosis results and do not require prior expertise, are becoming more and more popular nowadays [9–14]. Generally, data-driven techniques for fault diagnosis are carried out under the assumption that training and testing data are subject to the same distribution. However, in real-world applications, due to variations of environment, operating condition, bearing quality *etc.*, the distributions of training and testing data are usually different from each other, that deteriorates the generalization ability of applying the pattern knowledge learned from the labeled training data, denoted as *source domain*, to the new unlabeled testing data, denoted as *target domain*. This challenge of pattern learning validity is known as the domain shift problem [15].

Figure 1 presents an illustration of domain shift. While the classifier can be effectively trained using the labeled source domain data, it loses the classification validity on the target domain due to the existence of domain shift. That leads to serious performance degradation in fault diagnosis. This paper proposes a novel deep learning method for rolling bearing fault diagnosis using multi-layer domain adaptation. As shown in Figure 1, the domain shift problem is expected to be solved by jointly minimizing the classification error and the distribution discrepancy between the source and target domains.

Traditionally, many signal processing methods have been applied to machinery fault signal analysis, including wavelet analysis [2, 3], stochastic resonance techniques [4, 16, 17] and so forth [5–8]. In the past decade, a large number of studies have been carried out based on machine learning and statistical inference techniques, such as artificial neural networks (ANN) [9, 10, 18], support vector machines (SVM) [13, 14], random forest (RF) [19], fuzzy inference and other improved algorithms [11, 12]. In general, neural networks are one of the most popular data-driven methods to identify faulty and healthy machine conditions. Fault diagnosis is treated as a classification problem through feature extraction. First, raw input signal is mapped into representative features. The health condition and the corresponding fault location and severity are identified according to the extracted features afterwards.

Recently, deep learning network is emerging as a highly effective network structure for pattern recognition, that holds the potential to overcome the obstacles in the current intelligent fault diagnosis. Deep learning is characterized by the deep network architecture where multiple layers are stacked in the network to fully capture the representative information from raw input data [20]. High-level abstractions of data can be modeled well with the help of the complex deep structures, leading to more efficient feature extraction compared with the shallow networks. Deep learning methods have gained great interests and achieved significant results in machinery fault diagnosis researches [21–27]. In this study, deep learning is used as the main architecture for fault diagnosis.

Domain Adaptation (DA) is a particular case of transfer learning that leverages labeled data in the source domain, to learn a classifier for unlabeled data in the target domain [15]. In the recent years, domain adaptation methods have been successfully developed and applied in many practical tasks such as sentiment analysis [28], object recognition in different situations [29, 30], facial recognition [31], speech recognition [32], video recognition [33] *etc.* Generally, it is assumed that the task is the same for different domains, i.e. class labels are shared, and the source domain is related to the target domain. However, the two domains are not subject to the same distribution. The domain discrepancy poses an obstacle in adapting the well-trained models across domains. In bearing fault diagnosis, domain shift is very common in industries. For instance, with respect to the same fault location and severity classification task, the distributions of the data are possible to differ significantly with different rotating speed and motor load. Basically, applying the learned fault patterns on new operating conditions requires specific customization to accommodate the new domain data. One solution is by the means of acquiring a certain number of valid and labeled data in the target domain. However, that is time consuming and expensive in most cases, and even not feasible in some practical applications. On the other hand, the labeled source domain data and unlabeled target domain data can be further explored to calibrate the established model in order to achieve promising performance in the new situations, that is relatively easy to implement and preferred in real-world applications. This approach can be achieved by either adapting the established model trained from the source domain using the unlabeled target data, or developing a new model with the all the available data.

Domain adaptation establishes knowledge transfer from the source domain to the target domain by exploring domain-invariant structures that bridge the distribution discrepancy [34]. In the past years, a large number of researchers have been trying to build the domain-invariant model from data, which minimizes the distribution discrepancy in the latent feature space. In [35–37], shallow domain-invariant features are learned by minimizing the discrepancy. Furthermore, latest researches have revealed that deep learning architectures for domain adaptation are able to learn more transferable features and thus are more promising [38, 39].

In general, the deep architecture extracts features from generic to task-specific ones through the layers. Some studies find that the feature transferability drops significantly in the higher layers with increasing domain discrepancy [38], while others report that the lower layers may be more responsible for the domain biases [40]. Based on the latest understanding of domain discrepancy in the literature, we attempt to enhance the feature transferability by minimizing the distribution discrepancy throughout the deep network. Specifically, the representations of multiple layers are embedded to a reproducing kernel Hilbert space where the mean embeddings of different domain distributions can be explicitly matched. Since the mean embedding matching is inevitably influenced by the kernel selections, a multi-kernel approach is further designed to leverage different kernels and formulate a principled approach for optimal kernel selection.

Despite the success achieved by domain adaptation, limited researches can be found with respect to its application on fault diagnosis. Lu *et al.* proposed a deep neural network-based domain adaptation method for diagnosis, where the feature maximum mean discrepancy (MMD) is minimized, and a weight regularization term is used to strengthen the representative features. An adaptive batch normalization method was proposed by Zhang and colleagues [41] to improve the domain adaptation ability of neural network. Xie *et al.* [42] addressed the cross-domain feature extraction and fusion from time and frequency-domain with spectrum envelop pre-processing and time domain synchronization average principle using transfer component analysis (TCA). The source domain data are used as auxiliary data to assist target data classification in [43].

This paper proposes a novel deep convolutional neural network-based domain adaptation method for rolling bearing fault diagnosis. Machinery vibration data are used as model inputs. Labeled source domain data and unlabeled target domain data are assumed to be available. Different from existing researches, multi-layer and multi-kernel maximum mean discrepancies between the source and target domain data are minimized to address the domain shift problem. Experiments on a popular rolling bearing dataset are carried out to validate the effectiveness of the proposed method. The diagnosis performance is extensively evaluated in different scenarios. It is illustrated that multiple layers, rather than only the last ones, contribute to the domain biases through the network, and the necessity of the application of multi-layer MMD is presented. The superiority of the proposed method is demonstrated by comparing with other approaches and related works using the same dataset.

The remainder of this paper starts with the theoretical background in Section 2. The domain adaptation problem, convolutional neural network, MMD and softmax classifier are introduced. The proposed fault diagnosis method is presented in Section 3, and experimentally validated using a popular rolling bearing dataset in Section 4. We close the paper with conclusions in Section 5.

2 Theoretical Background

In this section, the domain adaptation problem for fault diagnosis is formulated, and the preliminaries for convolutional neural network, maximum mean discrepancy and softmax regression are introduced.

2.1 Domain Adaptation Problem

Traditionally, machinery fault diagnosis aims to identify fault location, severity *etc.* based on a prior known set of faults. It is assumed that the source and target domain distributions are the same, and the learned fault patterns from the labeled training samples can be directly applied on the unlabeled testing samples. However, discrepancy between the source and target domains inevitably exists in practical tasks, which makes the model generalization ability deteriorate across domains.

In this paper, we focus on the domain adaptation methods in order to better generalize the learned fault patterns from the source domain to the target domain. In general, this study is carried out under the assumptions:

(1) The fault diagnosis task remains the same for different domains, i.e. the class labels are shared.
(2) The source and target domains are related to each other, but have different distributions.
(3) Labeled samples from the source domain are available for training.
(4) Unlabeled samples from the target domain are available for training and testing.

In domain adaptation problems with respect to fault diagnosis, let X denote the input space and $Y = \{1, 2, ..., N_c\}$ represents the set of N_c possible machine health conditions. We are given a source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ of n_s labeled samples and a target domain $\mathcal{D}_t = \{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$ of n_t unlabeled samples. $\mathcal{D}_s$ and $\mathcal{D}_t$ are sampled from joint distributions $P(X, Y)$ and $Q(X, Y)$ respectively, and $P \neq Q$. The purpose of this paper is to construct a deep neural network $\mathbf{y} = f(\mathbf{x})$ that is able to reduce the cross-domain shifts in joint distributions and learn domain-invariant features and classifiers, in order to minimize the target risk $R_t(f) = \Pr_{(\mathbf{x},\mathbf{y})\sim Q}[f(\mathbf{x}) \neq \mathbf{y}]$ with source supervision.

2.2 Maximum Mean Discrepancy

In this study, the maximum mean discrepancy (MMD) is adopted to measure the discrepancy between distributions [44]. MMD is defined as the squared distance between the kernel embeddings of marginal distributions in the reproducing kernel Hilbert space (RKHS).

$$\mathrm{MMD}_k(P, Q) \triangleq \left\| \mathbf{E}_P[\phi(\mathbf{x}^s)] - \mathbf{E}_Q\left[\phi\left(\mathbf{x}^t\right)\right] \right\|_{\mathcal{H}_k}^2, \tag{1}$$

where $\mathcal{H}_k$ denotes the RKHS endowed with a characteristic kernel k. The most important property is $\mathrm{MMD}_k(P, Q) = 0$ iff $P = Q$.

As stated in [45], kernel choice is critical to ensure the testing power and low testing error of MMD, since different kernels may embed probability distributions in different RKHSs where different orders of sufficient statistics can be emphasized. Therefore in this study, we adopt multiple kernels of MMD to leverage different kernels and formulate a principled approach for optimal kernel selection. Specifically, a mixture of N_k RBF kernels are utilized,

$$k(\mathbf{x}^s, \mathbf{x}^t) = \sum_{i=1}^{N_k} k_{\sigma_i}(\mathbf{x}^s, \mathbf{x}^t), \tag{2}$$

where k_{σ_i} represents a Gaussian kernel with bandwidth parameter σ_i. In the experiments, it is found that using simple values of the bandwidth prameters and a mixture of 5 kernels is able to obtain good results [46]. Therefore, the default bandwidth parameters are selected as 1, 2, 4, 8 and 16 in this study, and their weights are kept equal for simplicity.

2.3 Convolutional Neural Network

Convolutional neural networks (CNNs), that are specifically designed for variable and complex signals, are utilized in this study. In the past few years, a large number of researches [24–27] have benefited from CNN's characteristics of local receptive fields, shared weights and spatial sub-sampling. CNN's ability to maintain data information regardless of scale, shift and distortion invariance has been shown [47].

The convolutional layers convolve multiple filters with raw input data and generate features, and the following pooling layers extract the most significant local features afterwards. The input data are usually 2-dimensional (2D) data for CNNs to learn abstract spatial features by alternating and stacking convolutional kernels and pooling operation. Since the input data in this research is a sequence of machinery vibration signal, the 1-dimensional (1D) CNN is briefly introduced in the following.

The input sequential data is assumed to be $\mathbf{x} = [x_1, x_2, ..., x_N]$ where N denotes the length of the sequence. The convolution operation in the convolutional layer can be defined as a multiply operation between a filter kernel $\mathbf{w}$, $\mathbf{w} \in R^{F_L}$, and a concatenation vector representation $\mathbf{x}_{i:i+F_L-1}$, which can be expressed as,

$$\mathbf{x}_{i:i+F_L-1} = x_i \oplus x_{i+1} \oplus \cdots \oplus \ x_{i+F_L-1}, \tag{3}$$

where $x_{i:i+F_L-1}$ represents a window of F_L length sequential signal starting from the i-th point, and $\oplus$ concatenates the data samples into a longer embedding. The final convolution operation is defined as,

$$z_i = \varphi(\mathbf{w}^T \mathbf{x}_{i:i+F_L-1} + b), \tag{4}$$

where $*^T$ denotes the transpose of a matrix $*$, and b and φ represent the bias term and non-linear activation function, respectively. The output z_i can be considered as the learned feature of the filter kernel $\mathbf{w}$ on the corresponding subsequence $\mathbf{x}_{i:i+F_L-1}$. By sliding the filter window from the first point to the last point in the sample data, the feature map of the j-th filter can be obtained, which is denoted as,

$$\mathbf{z}_j = [z_j^1, z_j^2, ..., z_j^{N-F_L+1}]. \tag{5}$$

In CNNs, multiple filter kernels can be applied in the convolution layer with different filter length F_L.

Usually, a pooling layer is applied to the feature maps generated by the convolutional layer. On the one hand, the pooling is able to extract the most significant local information in each feature map. On the other hand, the feature dimensionality, i.e. the number of model parameters, can be remarkably reduced by this operation. In general, average-pooling and max-pooling are widely used. In this paper, the max-pooling function is applied in the network.

The max-pooling operation is carried out in the feature maps with a pooling length of g. The extracted feature corresponding to the filter kernel can be obtained as,

$$\mathbf{p}_j = [p_j^1, p_j^2, ..., p_j^s], \tag{6}$$

$$p_j^k = \max\left(z_j^{(k-1)g+1}, z_j^{(k-1)g+2}, ..., z_j^{kg}\right), \tag{7}$$

where $\mathbf{p}_j$ is the output of the pooling layer applied to the j-th feature map and has s dimensions. By alternating the convolutional and max-pooling layers, fully-connected layer and softmax regression are usually added as the top layers to make classification. The framework for 1D CNN is displayed in Figure 2.

2.4 Softmax Classifier

A softmax regression is usually implemented on the top layer in a neural network for multi-class classification [48]. The information derived from multiple hidden layers is used as input of a supervised classifier followed by global back-propagation optimizations. In this paper, softmax regression is employed as the machinery health condition classifier in the network.

The training samples are denoted as $\mathbf{x}^{(i)}$ and the corresponding label set is $y^{(i)}$ where $i = 1, 2, ..., N_{train}$ and N_{train} is the number of the training samples. $\mathbf{x}^{(i)} \in R^{N\times 1}$ and $y^{(i)} \in \{1, 2, ..., K\}$ where K is the number of the labeled categories. For an input sample $\mathbf{x}^{(i)}$, the softmax regression is able to estimate the probability $p(y^{(i)} = j \mid \mathbf{x}^{(i)})$ for each label j ($j = 1, 2, ...K$). The estimated probabilities of the input data $\mathbf{x}^{(i)}$ belonging to each label can be obtained according to the hypothesis function,

$$J_\theta(\mathbf{x}^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \theta) \\ p(y^{(i)} = 2 \mid \mathbf{x}^{(i)}; \theta) \\ \vdots \\ p(y^{(i)} = K \mid \mathbf{x}^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{k=1}^{K} e^{\theta_k^T \mathbf{x}^{(i)}}} \begin{bmatrix} e^{\theta_1^T \mathbf{x}^{(i)}} \\ e^{\theta_2^T \mathbf{x}^{(i)}} \\ \vdots \\ e^{\theta_K^T \mathbf{x}^{(i)}} \end{bmatrix}, \tag{8}$$

where $\theta = [\theta_1, \theta_2, ..., \theta_K]^T$ denotes the softmax model parameters. This classifier makes sure the outputs are positive and sum to 1, allowing us to interpret the outputs of the network as the probabilities for each class.

3 Proposed Fault Diagnosis Method

3.1 Network Architecture

In this paper, we focus on the investigation of domain adaptation, and a conventional deep CNN architecture is used for simplicity. Figure 3 shows the structure of the proposed network for machinery fault diagnosis. In general, the proposed deep learning method combines two architectural ideas for better feature extraction of vibration signals, i.e. CNN and fully-connected layer.

First, four stacked 1D convolutional layers are designed for feature extraction, and they are supposed to share the same configuration for convenience. F_N local filters of F_L length window size are used in each convolutional layer, and zeros-padding operation is implemented to keep the feature map dimension from changing [49]. Each convolutional layer is followed by a max-pooling layer to reduce the data dimension while keeping the significant spatial information. In this way, different levels of representations of input data are obtained.

Next, the learned high-level feature representations are flattened and connected to a fully-connected (FC) layer. Dropout technique is used in this layer with rate of 0.5 to avoid overfitting [22]. Finally, a softmax regression is adopted to predict the fault categories.

Batch normalization (BN) is able to accelerate the training process especially for deep network and has achieved good performance in deep learning recently [50]. In this study, BN is used after each convolution layer and before activation. In addition, the rectified linear units (ReLU) activation functions are generally used in the network [51]. They do not suffer from gradient vanishing or gradient diffusion in the training process. Therefore, better performance can be usually achieved especially in deep architecture [52].

Based on previous researches, the network performance can be significantly affected by the number and size of the convolutional filters. Since the convolutional parameters are closely related to the distribution discrepancy of the layers, they are expected to have remarkable influence on the proposed method, and that will be investigated in Section 4.3.4. By default, the filter number is 10 and the filter length is 10 in this paper, and the fully-connected layer has 256 neurons for the final regression.

3.2 Optimization Objective

As pointed out in Section 2.1, the fault diagnosis task is the same for different domains in this study, that indicates the class categories are shared. Since labeled training samples are available, the first optimization objective is to minimize the classification error of the training samples. In this case, the cross-entropy function $\mathcal{L}_c$ is used as the loss function [53].

Corresponding with the softmax classifier in Section 2.4, $\mathcal{L}_c$ is defined as,

$$\mathcal{L}_c = -\frac{1}{N_{train}} \left[\sum_{i=1}^{N_{train}} \sum_{k=1}^{K} 1 \left\{ y^i = k \right\} \log \frac{e^{\theta_k^T \mathbf{x}^{(i)}}}{\sum_{j=1}^{K} e^{\theta_j^T \mathbf{x}^{(i)}}} \right]. \tag{9}$$

Furthermore, in order to effectively generalize the classifier trained from the source domain to the target domain, the two domain distributions of the input data are supposed to be drawn closer to each other from the learned feature representations, i.e. the hidden layers in the network. Therefore, the multi-kernel MMD between distributions as described in Section 2.2 is used as the optimization objective.

While MMD has been adopted in the literature and achieved good domain adaptation results, most researches aim to minimize the distribution discrepancy in the last layer of the network. However, as pointed out in [34], feature transferability deteriorates in multiple top layers, and adapting a single layer is not able to effectively eliminate the bias between the source and target domains since other layers may not be transferable. Moreover, recent researches pointed out that the first layers are susceptible to domain shift even more than the later layers [40]. Therefore in this paper, we develop the domain adaptation method by jointly adapting the representation layers and the classifier, and the multiple layer MMD is adopted. That is expected to bridge the domain discrepancy underlying both the marginal distribution and the conditional distribution [54]. The MMD loss is defined as,

$$\mathcal{L}_m = \sum_{l \in \mathbf{L}} \mathrm{MMD}_k(P^l, Q^l), \tag{10}$$

where $\mathbf{L}$ denotes the layers between which the MMD loss is computed and used, P^l and Q^l are the l-th layer representations for the source and target samples respectively, and $\mathrm{MMD}_k(P^l, Q^l)$ represents the multi-kernel MMD between the source and target domains evaluated on the l-th layer representation.

By integrating Equations (9) and (10), the final objective function can be expressed as,

$$\mathcal{L}_{final} = \mathcal{L}_c + \alpha \mathcal{L}_m, \tag{11}$$

where α denotes the penalty coefficient whose default value is 1 in this paper.

3.3 Diagnosis Procedure

The flow chart of the proposed fault diagnosis method is presented in Figure 5. First, the raw machinery vibration signals are collected by sensors, and the training and testing samples are prepared based on the labeled source domain and unlabeled target domain data. The raw vibration data are used directly as the model input. No hand-crafted signal processing feature is needed, such as skewness, kurtosis *etc.* Therefore, no prior expertise on fault diagnosis is required in the proposed method, that facilitates the industrial applications.

Next, based on the specific fault diagnosis problem and the dataset information, the network configuration is determined. In this study, since we focus on domain adaptation problem, a conventional deep CNN architecture is used as described in Section 3.1. The *Xavier* normal initializer is employed for the initializations of the network weights and biases [55].

To start the learning process, both the labeled source domain data and the unlabeled target domain data are fed into the proposed network. Domain-invariant features of the vibration signals are extracted through the multiple convolutional and pooling layers. The distribution discrepancies in multiple layers are minimized. On the top of the network, the fully-connected layer and softmax regression are applied to classify the rolling bearing health conditions with the learned domain-invariant features from the deep network. The back-propagation (BP) algorithm [56] is applied for the updates of all the parameters in the network, and the Adam optimization method [57] is used to minimize the objective (Equation (11)) with the whole batch. After training for 2000 epochs, the loss of the proposed network converges in general. Therefore, the number of training epochs is 2000 by default.

The testing samples from the target domain will be fed into the proposed network when the training process is over, and the testing fault diagnosis results can be obtained.

4 Experimental Study

4.1 Experimental Setup

The rolling bearing dataset used in this study is provided by the Bearing Data Center of Case Western Reserve University [58]. The dataset is composed of multi-variate vibration signals generated by a bearing test-rig, as presented in Figure 4.

The main components of the experimental apparatus are a 2-horsepower (hp) motor (left side of figure), a torque transducer/encoder (center of figure) and a dynamometer (right side of figure). The motor shaft is supported by 6205-2RS JEM SKF bearings. These bearing data are collected by acceleration transducers under four load conditions (0, 1, 2 and 3 hp) with sampling rates of 12kHz. The motor rotational speed varies between 1730 and 1797 rpm depending on the load.

The vibration signals used in this study were collected from the drive end of the motor in the test rig on four different health conditions: 1) normal condition (H); 2) outer race fault (OF); 3) inner race fault (IF); and 4) ball fault (BF). All the three kinds of faults are generated by electro-discharge machining with fault diameters of 7 mils, 14 mils and 21 mils (1 mil = 0.001 inches), respectively. Therefore, this dataset contains 10 bearing health conditions under the four loads, where the same health condition under different loads is treated as 1 class.

In this paper, the proposed method is evaluated across 6 transfer tasks, i.e. T_{01}, T_{02}, T_{03},

T_{30}, T_{31} and T_{32}. The subscripts are intuitive to understand. For instance, the task name T_{01} denotes the scenario where the labeled data with 0 hp load are considered as the source domain for supervised training, and the unlabeled data with 1 hp load are the target domain for testing. The 6 tasks provide a general evaluation of the proposed method with respect to the model transferability in different operating situations.

N_{sou} and N_{tar} samples for each health condition under one load are supposed to be selected as the source and target domain data, respectively. In the case studies of this paper, we assume $N_{sou} = N_{tar}$ in the training process for simplicity. After the networks finish training, N_{test} samples of each class are selected for cross-domain testing. In order to fully examine the validity of the fault diagnosis methods, $N_{test} = 400$ is used as the default value which indicates $400 \times 10 = 4000$ samples are tested.

For the convenience of classification, the 10 health conditions with different fault location and fault size are artificially set as class label 1 to 10, respectively. The detailed information of the dataset and the transfer tasks is presented in Tables 2 and 3.

For each raw collected signal sequence that represents one working condition, the first 120000 points are used for selecting samples. The raw data sequence is equally divided into N_{sou} or N_{tar} sub-signals based on the specific task and each sub-signal contains N_{input} sequential points. In this study, different amount of data are used to evaluate the proposed method, that will be presented in Section 4.3. By default, $N_{input} = 500$ is used, and data overlapping is generally avoided in the sampling process.

4.2 Comparison Approaches

In order to present a complete evaluation of the proposed method, different implementations are carried out for comparison. The latest related researches on the same dataset are also presented to show the effectiveness and superiority of the proposed method. Specifically, the following approaches are studied.

(1) ML-MK-I-(**x**)

The abbreviations denote the proposed fault diagnosis methods where the **M**ulti-**L**ayer and **M**ulti-**K**ernel MMD, and the **I**ntegrated optimization objective as expressed in Equation (11) are used. **x** represents the layers whose feature MMD are minimized as Equation (10) shows. For instance, the proposed method ML-MK-I-(1,2,3,4,fc) denotes the situation where the MMDs of the convolutional layers with number **1**, **2**, **3** and **4**, and the **f**ully-**c**onnected layer are included in the optimization objective. In the following sections, different layer combinations of **x** are evaluated for comparisons in various situations.

(2) MK-I

Most existing researches on unsupervised domain adaptation focus on minimizing the distribution discrepancy in the last layer of high-level extracted features. In order to show the superiority of the proposed multi-layer MMD method, the MK-I approach, which is equivalent to ML-MK-I-(fc), is implemented for comparison where only the MMD in the final fully-connected layer is considered.

(3) ML-I-($\mathbf{x}$)

ML-I-($\mathbf{x}$) is examined as comparison to show the performance improvements by the multi-kernel MMD. In this study, ML-I-(1,2,3,4,fc) which corresponds with the proposed method is implemented for comparison, and the one MMD kernel with bandwidth parameter of 4 is used.

(4) T-S

Instead of training a generic network with both the source and target domain data, some researches [59] suggest performing adaptation by learning a **T**arget-**S**pecific network from the source-specific network. That is an alternative solution for the domain adaptation problem and is evaluated in this paper.

First, a network is trained with the labeled source domain data, which is denoted as the source-specific network. Next, a new similar network, i.e. the target-specific network, is built and trained with the unlabeled target domain data. The layer distribution discrepancies between the two networks are minimized as the optimization objective for the new network. Finally, the high-level features extracted by the target-specific network are classified by the source-specific classifier for fault diagnosis. In this way, the domain-invariant features are also expected to be learned.

(5) Two-Stage Learning

As proposed in recent studies [60], a two-stage learning method for domain adaptation can be adopted, where the integrated objective in Equation (11) is separated. Specifically, after initialization with the labeled source domain data, the generic network can be further trained with both the source and target domain data by minimizing the distribution discrepancies through layers (Equation (10)). Afterwards, the classifier is determined using the labeled source domain data only (Equation (9)).

However, the network weights have high probability of dropping to near-zeros without proper regularizations using this method [60]. The regularizations do not offer noticeable improvements to the proposed method, and thus there is not a fair basis to compare with the two-stage learning method in this paper. Diagnosis performance of that method under similar experimental settings can be found in [60].

(6) Without-DA

At last, the traditional training method without domain adaptation (DA) is implemented for comparison. In this case, only the Equation (9) is used as the optimization objective.

4.3 Experimental Results and Performance Analysis

In this section, the cross-domain fault diagnosis results of the proposed method on the rolling bearing dataset are presented, as well as comparisons with other approaches and related researches on the same dataset. In this paper, the reported experimental results are averaged by 10 trials to reduce the effect of randomness, and the mean values and standard deviations are provided. All the experiments are carried out on a PC with Intel Core i7

CPU, 8-GB RAM and GEFORCE GTX 950M GPU. *Tensorflow* platform is used for the programming, and GPU parallel computing is employed to accelerate the computing.

4.3.1 Results on the Rolling Bearing Dataset

4.3.1.1 Transfer Task $\mathbf{T}_{03}$ Figure 6 shows the comprehensive experimental results of the cross-domain fault diagnosis task T_{03}. In order to better illustrate the performance of different methods, different amount of the labeled source domain data (N_{sou}) are used for training. The compared methods are implemented with the default experimental setting in this section.

First, it is noted that training with more labeled source domain data leads to higher testing accuracy for all the methods. That is consistent with the related studies on deep learning that sufficient training samples are usually required for good network performance, and the cross-domain tasks also follow this pattern.

Furthermore, it can be observed that generally, significant improvements in cross-domain diagnosis can be achieved by domain adaptation. Basically, low testing accuracies are obtained using the Without-DA method in all the cases, while for the rest of the methods with domain adaptation, remarkable increases in the testing accuracies are obtained.

Specifically, the proposed method achieves the best diagnosis performance in all the scenarios. The testing accuracies achieved by the ML-I-(1,2,3,4,fc) method are generally smaller than those of the proposed method, that shows the improvements by the proposed multi-kernel MMD. The diagnosis approaches with a single layer MMD minimization are also evaluated, i.e. ML-MK-I-(1), ML-MK-I-(2), ML-MK-I-(3), ML-MK-I-(4), and MK-I (ML-MK-I-(fc)). While the MK-I method which minimizes the distribution discrepancy in the final fully-connected layer provides relatively good results, the rest approaches achieve limited improvements comparing with the Without-DA method. As variations of the proposed multi-layer MMD method, ML-MK-I-(1,2,fc) and ML-MK-I-(3,4,fc) are also evaluated, and they have obtained satisfactory results, which are slightly worse than the proposed method, but obviously better than the rest. Therefore, domain adaptation with multiple layer MMD is well suited for the cross-domain problems, and it is preferred to consider the distribution discrepancies in all the layers in this case study.

Moreover, the T-S method which provides an alternative way for domain adaptation, achieves good testing results with large labeled training dataset. However, the diagnosis performance deteriorates significantly when N_{sou} becomes smaller. That is because for the T-S method, the target-specific network is trying to minimize the distribution discrepancy between the target domain data and the learned representations from the source-specific network. In this way, the features which are learned for the source domain data are enhanced, rather than the domain-invariant features directly learned by the proposed network.

The presented experimental results demonstrate the effectiveness and superiority of the proposed method. Especially, comparing with other approaches, the improvements by the proposed method are more significant with small training dataset. Therefore, the proposed method has large potential for industrial applications since the valid and labeled training data are always difficult to obtain in practice.

4.3.1.2 The Other five Transfer Tasks The experimental results of the other 5 transfer tasks, i.e. T_{01}, T_{02}, T_{30}, T_{31} and T_{32}, are presented in Figures 7 and 8, where different amount of the labeled source domain data are used for training. In general, the display patterns of the diagnosis performance using different methods are similar with those in Figure 6. The testing accuracies in some tasks such as T_{01} and T_{02} are slightly higher than that of T_{03} as presented in Figure 6. That is due to the fact that the related source and target domains are closer to each other by nature. For instance, the difference in motor load between 0 hp and 1 hp (T_{01}) is smaller than that between 0 hp and 3 hp (T_{03}), which makes it easier to transfer the learned representations from 0 hp domain to 1 hp domain. Moreover, the high accuracies in the tasks T_{30}, T_{31} and T_{32} indicate the proposed method performs well bidirectionally between domains.

In addition, corresponding with the findings from Figure 6, larger cross-domain diagnosis improvements can be achieved by the proposed method comparing with other approaches when small labeled source domain dataset is used. Therefore, the proposed method is able to achieve the best cross-domain diagnosis performance in different transfer tasks, and its effectiveness and robustness are further validated.

4.3.2 Visualization of Learned Representation

In this section, the effectiveness of the proposed method for fault diagnosis is illustrated qualitatively based on visualization of learned representation. Since the last hidden layer in the network, i.e. the fully-connected layer, is directly responsible for the final fault classification, and it is generally agreed in the literature that the last hidden layer is of great importance in cross-domain problems, the visualization of the fully-connected layer is presented in this section.

An effective technique “t-SNE” is used to visualize the high-dimensional data representation by mapping the samples from the original feature space into a 2-dimensional space map [61]. The principal component analysis (PCA) is first adopted to reduce the dimensionality of the feature data to 50 and suppress signal noise. Afterwards, the technique “t-SNE” is used to convert the 50-dimensional learned representation to a 2-dimensional map.

Take the task T_{03} for instance, Figure 9 shows the resulting maps of the learned representations in the fully-connected layer for both the source and target domains. The visualizations of the features using the proposed method (ML-MK-I-(1,2,3,4,fc)) and those without domain adaptation (Without-DA) are both presented.

It can be seen that the features extracted by the proposed method cluster the best where all the data samples of different conditions are separated well. That is the basic requirement for high fault diagnosis accuracy. Furthermore, good fusion of the source and target domains is observed. The samples in the two domains that belong to the same fault class mostly cluster into the same region, and only a small amount of cross-region data overlappings are observed. For each class, the samples from the source and target domains practically overlap together, that facilitates the final regression for the cross-domain classification.

On the other hand, for the Without-DA method without domain adaptation, while the samples with the same health condition labels cluster well using the same network architecture with the proposed method, significant separations between the source and target domains are observed. For most of the fault classes, the two domains of the same class are not projected into the same region, and the target domain samples merge into the regions of other classes. Note that for the well-trained network, the feature space is divided into multiple regions corresponding with different labels by the softmax classifier. Distribution discrepancy between the source and target domains in the fully-connected layer directly leads to worse testing classification results. In this way, the necessity of domain adaptation is demonstrated.

It should be pointed out that the final classification is carried out in a high-dimensional space nonlinearly. Therefore, acceptable point overlappings for different health conditions in visualization agree with the high classification accuracies presented in Figure 6.

4.3.3 Layer Distribution

In this section, comprehensive distribution discrepancies between the source and target domains in each layer are illustrated, in order to show the improvements by the proposed multi-layer MMD method. The visualization approach used in this section is the same with that presented in Section 4.3.2. The experimental results of the transfer task T_{03} are adopted for demonstration, and $N_{sou} = 100$ is used.

Figure 10 shows the visualizations of the two domains in each layer of the proposed network. Three methods are compared, i.e. Without-DA, ML-MK-I-(1) and ML-MK-I-(1,2,3,4,fc). It can be clearly observed that generally, distribution discrepancy between domains exists in all the layers in the network. Especially, the domain shift phenomenon is more obvious in the higher layers than in the lower ones. That can be explained that the lower layers are basically responsible for extracting generic features from signals, and the high-level abstract features that are task-specific are extracted by the higher layers.

Corresponding with the results presented in Figure 9, the source and target domains are drawn closer to each other in every layer using the proposed multi-layer MMD approach, that indicates the domain-invariant features rather than the task-specific ones are mostly extracted throughout the network. Therefore, the domain shift problem can be effectively solved with the proposed method.

In addition, as stated in many researches, since the lower-level filters are mostly generic, they can be considered domain-invariant themselves, and domain adaptation can be thus focused on the higher layers. In order to show the effects of lower layers on domain adaptation, the layer visualization of the ML-MK-I-(1) method is presented in the middle column in Figure 9. In this case, only the distribution discrepancy in the first convolutional layer is minimized. It is observed that the general distances between the two domains using ML-MK-I-(1) are noticeably smaller than those without domain adaptation for all the layers. That indicates the domain-invariant information generated in the first layer propagates forward to the higher layers. While the first layers are usually considered generic, i.e. they extract features regardless of the dataset, they still have large potential to further relieve the domain shift problem. Therefore, the representative information in the first layers also conveys the dataset characteristics, and is supposed to be utilized for better domain adaptation.

Furthermore, the distribution MMD in each layer with different methods are exhibited in Figure 11. It is clearly observed that the first layers can be considered generic to some extent, and generally have low MMD even for the Without-DA method. For the higher layers especially the fully-connected layer, the differences in MMD become significant. It is found that generally the MMD in the fully-connected layer using different method is inversely proportional to the testing diagnosis accuracy presented in Figure 6. That is consistent with the fact that the fully-connected layer is directly related to the final classification, and larger discrepancy leads to worse cross-domain diagnosis performance. Figure 11 also shows the superiority of the proposed method, which is able to achieve the smallest MMD in each layer in general, resulting in the highest cross-domain classification accuracy. Additionally, since the ML-I-(1,2,3,4,fc) method uses only one kernel of MMD, its corresponding metric is relatively smaller than others with five kernels, and its results are thus not necessarily better than the proposed method.

4.3.4 Effects of Parameters

In the proposed fault diagnosis method, the convolutional filter number and size are two main parameters that affect the network performance. The introduced punishment factor α, which determines the domain adaptation strength, may also have influence on the diagnosis accuracy. In this section, we discuss the effects of the associated parameters. The experimental setting is similar with that in previous sections, and the transfer task T_{03} is used for illustration. In order to better present the difference in diagnosis performance by different parameters, $N_{sou} = 20$ is adopted in this case study.

Figure 12 shows the impacts of the convolutional filter size and number on the testing diagnosis accuracy. In general, significant improvements can be achieved by larger F_N and F_L. More convolutional filters in each layer and larger filter size contribute to the learning capacity of the deep neural network. Additionally, the long distance information in the sequential vibration signal can not be effectively captured with small window size in some cases [25].

Specifically, it is observed that the influence of the filter number is more remarkable than that of the filter size. As F_N becomes larger, the average testing accuracy increases stably, and it reaches 99.75% when 50 filters are adopted in each layer. On the other hand, the improvements by larger filter size are relatively limited, and $F_L = 50$ leads to 96.65% average accuracy in this case.

Therefore, generally large values of F_N and F_L are suggested to improve the cross-domain diagnosis performance. However, as shown in Figure 12, the computing time for network training becomes longer with larger F_N and F_L. A tradeoff is supposed to be made between the diagnosis accuracy and the computational burden. Since the network training process is implemented off-line, the longest average computing time of 865.5 seconds for 2000 epochs in this case is still acceptable in the proposed fault diagnosis framework.

Another important coefficient that affects the network performance is the sample dimension N_{input}. Based on previous studies and related works, larger N_{input} generally leads to better diagnosis performance. However, the variation of N_{input} with the same number of training samples indicates the training dataset changes implicitly. Therefore, it is difficult to provide a fair comparison basis to study the impacts of N_{input}.

In addition, experiments are also carried out to investigate the influence of the introduced punishment factor α, whose default value is 1 in this paper. However, it is found that the testing diagnosis accuracy is not significantly affected by α in the case study, and the testing results generally keep stable with respect to different α in a reasonable range such as $[0.1, 10]$. That suggests the proposed method is robust to the selection of α.

4.3.5 Comparing with Related Works

The rolling bearing dataset used in this paper is very popular in machinery fault diagnosis researches, and many state-of-the-art classification results have been reported in the past years. However, very limited work can be found on cross-domain problem, and most studies focus on diagnosing bearing health condition using the training and testing data from the same domain.

In the latter case, 95% and higher testing accuracies were achieved in [62–64] where 4 bearing health conditions or fewer were considered. In [65], [66] and [67], 10 or more bearing conditions were classified, and 88.9%, 92.5% and 97.9% testing accuracies were obtained respectively. A two-stage machine learning method was proposed in [48] based on unsupervised feature learning and sparse filtering. Fairly high diagnosis accuracy of 99.66% was achieved.

On the other hand, domain adaptation problem was studied in [60], where the labeled data under 0 hp load were used as the source domain and the unlabeled data under 3 hp load were considered the target domain. That case study is similar with the T_{03} transfer task in this paper. Specifically, 4 health conditions were considered in [60], and 1000 samples of $N_{input} = 1200$ sample length were selected from both the source and target domains for training. As a result, as high as 94.73% cross-domain classification accuracy was achieved.

For the T_{03} transfer task in this study, using the default network configuration, the proposed method obtains the testing accuracy of 94.02% with 200 labeled source-domain samples ($N_{sou} = 20$), and 99.17% with 1000 labeled source-domain samples ($N_{sou} = 100$). Even higher testing accuracy is expected to be obtained using larger labeled training dataset. Furthermore, as presented in Section 4.3.4, the network configuration and the hyper-parameters have remarkable influence on the diagnosis performance. If the enhanced experimental settings of the proposed method is used regardless of the off-line computational burden for training, higher classification accuracy can also be achieved. For instance, using the parameters of $N_{sou} = 50$, $F_L = 20$ and $F_N = 50$, fairly high average testing accuracy of 99.76% is obtained, and the standard deviation is 0.17%. In fact, 100% testing accuracy is achieved in two out of ten trials, where all the 4000 target domain samples are precisely classified. Based on the experiments, even better results are obtained with the proposed cross-domain diagnosis method than those explicitly trained with the labeled target domain data in the literature. Considering 10 bearing health conditions are diagnosed in this paper, the proposed method is promising in solving domain shift problems.

In summary, the detailed comparison results with related works on the same rolling bearing dataset are presented in Table 4.

5 Conclusions

This paper proposes a novel deep learning-based machinery fault diagnosis method for domain adaptation. The maximum mean discrepancies between the source and target domains in multiple layers are minimized in order to solve the domain shift problem. Multiple kernels of MMD are used to leverage different kernels, and the MMD term is integrated with the classification loss for the network training. The effectiveness of the proposed method is validated by extensive experiments on a popular rolling bearing dataset. Comprehensive experiments are carried out to evaluate the robustness of the proposed method under different conditions, and assess the influence of the associated parameters on the diagnosis performance. Comparisons with other approaches and related researches on the same dataset are provided to verify the superiority of the proposed method.

Generally for data samples from two domains, while low transferring ability is observed for traditional neural networks, significant improvements in the cross-domain diagnosis performance can be achieved using domain adaptation techniques. Specifically in the case study of this paper, we attempt to diagnose the bearing health condition under a new motor load, that is different from the one under which the labeled data samples are available for training. Traditional methods fail to accurately predict the bearing fault class in the new domain. On the other hand, the domain adaptation methods are able to learn the domain-invariant features under the two different motor loads, and thus perform well diagnosing faults in the new domain.

It is observed in the experiments that the proposed method is well suited for the cross-domain fault diagnosis problem, and generally achieves the highest testing accuracy in different scenarios. The proposed method using multi-layer and multi-kernel MMD offers further improvements to the domain adaptation method, and outperforms the other comparing approaches. As high as 99.17% cross-domain testing accuracy is obtained with the default experimental setting, and up to 99.76% accuracy can be achieved using the enhanced network configuration.

Especially, the improvements by the proposed method on the cross-domain diagnosis performance are more significant when a smaller labeled training dataset is used. That indicates the overfitting problem, which is usually due to insufficient training data, can be relieved to some extent by the domain adaptation technique, and the proposed method has large potential for industrial applications since valid and precisely labeled data are always difficult to obtain in practice.

While satisfactory results have been obtained with the proposed method, the drawback lies in that currently the training dataset is balanced over different categories of bearing health conditions. In real applications, the bearing data in healthy condition (Class 1) are usually easy to obtain, while the data for different fault classes are scarce. Therefore, the next challenge is to efficiently extract the domain-invariant features between the source and target domains based on imbalanced training dataset. It is more difficult for transferring between imbalanced datasets than training on imbalanced dataset itself, since unlabeled data may significantly confuse the learned representations when the distribution is not balanced.

Moreover, as presented in the experimental results in this paper, the standard deviations of the testing results are not negligible in many scenarios. In the proposed framework, the initializations of the network weights and biases are mainly responsible for the deviations, which are supposed to be minimized for robustness. The two issues mentioned above will be focused on in further research, as well as the optimization of the hyper-parameters in the proposed method.

Acknowledgements

The material in this paper is based on work supported by grants (11172197, 11332008, and 11572215) from the National Science Foundation of China.

References

[1] H. Sun, Z. He, Y. Zi, J. Yuan, X. Wang, J. Chen, S. He, Multiwavelet transform and its applications in mechanical fault diagnosis - A review, Mechanical Systems and Signal Processing 43 (12) (2014) 1–24.

[2] P. W. Tse, Y. H. Peng, R. Yam, Wavelet analysis and envelope detection for rolling element bearing fault diagnosis - Their effectiveness and flexibilities, Journal of Vibration and Acoustics 123 (3) (2001) 303–310.

[3] Z. Ren, S. Zhou, C. E, M. Gong, B. Li, B. Wen, Crack fault diagnosis of rotor systems using wavelet transforms, Computers & Electrical Engineering 45 (2015) 33–41.

[4] X. H. Chen, G. Cheng, X. L. Shan, X. Hu, Q. Guo, H. G. Liu, Research of weak fault feature information extraction of planetary gear based on ensemble empirical mode decomposition and adaptive stochastic resonance, Measurement 73 (2015) 55–67.

[5] P. Zhou, S. Lu, F. Liu, Y. Liu, G. Li, J. Zhao, Novel synthetic index-based adaptive stochastic resonance method and its application in bearing fault diagnosis, Journal of Sound and Vibration 391 (2017) 194–210.

[6] G. He, K. Ding, H. Lin, Fault feature extraction of rolling element bearings using sparse representation, Journal of Sound and Vibration 366 (2016) 514–527.

[7] M. Žvokelj, S. Zupan, I. Prebil, EEMD-based multiscale ICA method for slewing bearing fault detection and diagnosis, Journal of Sound and Vibration 370 (2016) 394–423.

[8] S. Lu, X. Wang, Q. He, F. Liu, Y. Liu, Fault diagnosis of motor bearing with speed fluctuation via angular resampling of transient sound signals, Journal of Sound and Vibration 385 (2016) 16–32.

[9] G. F. Bin, J. J. Gao, X. J. Li, B. S. Dhillon, Early fault diagnosis of rotating machinery based on wavelet packets - Empirical mode decomposition feature extraction and neural network, Mechanical Systems and Signal Processing 27 (2012) 696–711.

[10] V. T. Tran, B. S. Yang, F. Gu, A. Ball, Thermal image enhancement using bi-dimensional empirical mode decomposition in combination with relevance vector machine for rotating machinery fault diagnosis, Mechanical Systems and Signal Processing 38 (2) (2013) 601–614.

[11] Z. Li, H. Fang, M. Huang, Diversified learning for continuous hidden Markov models with application to fault diagnosis, Expert Systems with Applications 42 (23) (2015) 9165–9173.

[12] A. Youssef, C. Delpha, D. Diallo, An optimal fault detection threshold for early detection using Kullback-Leibler divergence for unknown distribution data, Signal Processing 120 (2016) 266–279.

[13] X. Zhang, W. Chen, B. Wang, X. Chen, Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization, Neurocomputing 167 (2015) 260–279.

[14] R. Jegadeeshwaran, V. Sugumaran, Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines, Mechanical Systems and Signal Processing 5253 (2015) 436–446.

[15] G. Csurka, Domain adaptation for visual applications: A comprehensive survey, arXiv preprint arXiv:1702.05374 .

[16] H. L. He, T. Y. Wang, Y. G. Leng, Y. Zhang, Q. Li, Study on non-linear filter characteristic and engineering application of cascaded bistable stochastic resonance system, Mechanical Systems and Signal Processing 21 (7) (2007) 2740–2749.

[17] Q. Li, T. Wang, Y. Leng, W. Wang, G. Wang, Engineering signal processing based on adaptive step-changed stochastic resonance, Mechanical Systems and Signal Processing 21 (5) (2007) 2267–2279.

[18] B. Samanta, C. Nataraj, Use of particle swarm optimization for machinery fault detection, Engineering Applications of Artificial Intelligence 22 (2) (2009) 308–316.

[19] B. S. Yang, X. Di, T. Han, Random forests classifier for machine fault diagnosis, Journal of Mechanical Science and Technology 22 (9) (2008) 1716–1725.

[20] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504.

[21] C. Lu, Z. Y. Wang, W. L. Qin, J. Ma, Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification, Signal Processing 130 (2017) 377–388.

[22] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, X. Chen, A sparse auto-encoder-based deep neural network approach for induction motor faults classification, Measurement 89 (2016) 171–178.

[23] W. T. Mao, J. L. He, Y. Li, Y. J. Yan, Bearing fault diagnosis with auto-encoder extreme learning machine: A comparative study, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science (2016) 0954406216675896.

[24] X. Guo, L. Chen, C. Shen, Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis, Measurement 93 (2016) 490–502.

[25] W. Sun, R. Zhao, R. Yan, S. Shao, X. Chen, Convolutional discriminative feature learning for induction motor fault diagnosis, IEEE Transactions on Industrial Informatics 13 (3) (2017) 1350–1359.

[26] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, D. J. Inman, Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks, Journal of Sound and Vibration 388 (2017) 154–170.

[27] T. Ince, S. Kiranyaz, L. Eren, M. Askar, M. Gabbouj, Real-time motor fault detection by 1-D convolutional neural networks, IEEE Transactions on Industrial Electronics 63 (11) (2016) 7067–7075.

[28] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: A deep learning approach, 28th International Conference on Machine Learning (2011) 513–520.

[29] B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars, Unsupervised visual domain adaptation using subspace alignment, The IEEE International Conference on Computer Vision (ICCV) (2013) 2960–2967.

[30] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2066–2073.

[31] A. Sharma, D. W. Jacobs, Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 593–600.

[32] L. Samarakoon, K. C. Sim, On combining i-vectors and discriminative adaptation methods for unsupervised speaker normalization in DNN acoustic models, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 5275–5279.

[33] L. Duan, D. Xu, S. F. Chang, Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1338–1345.

[34] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: Proceedings of 32nd International Conference on Machine Learning, Vol. 37, 2015, pp. 97–105.

[35] X. Wang, J. Schneider, Flexible transfer learning under support and model shift, 27th International Conference on Neural Information Processing Systems (2014) 1898–1906.

[36] B. Gong, K. Grauman, F. Sha, Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation, 30th International Conference on Machine Learning 28 (2013) 222–230.

[37] S. J. Pan, I. W. Tsang, J. T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Transactions on Neural Networks 22 (2) (2011) 199–210.

[38] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, 27th International Conference on Neural Information Processing Systems (2014) 3320–3328.

[39] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: A deep convolutional activation feature for generic visual recognition, 31st International Conference on Machine Learning 32 (2014) 647–655.

[40] R. Aljundi, T. Tuytelaars, Lightweight unsupervised domain adaptation by convolutional filter reconstruction, in: Computer Vision - ECCV Workshops, Springer International Publishing, Cham, 2016, pp. 508–515.

[41] W. Zhang, G. Peng, C. Li, Y. Chen, Z. Zhang, A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals, Sensors 17 (2) (2017) 425.

[42] J. Xie, L. Zhang, L. Duan, J. Wang, On cross-domain feature fusion in gearbox fault diagnosis under various operating conditions based on transfer component analysis, in: Proceedings of IEEE International Conference on Prognostics and Health Management, 2016, pp. 1–6.

[43] F. Shen, C. Chen, R.-Q. Yan, R. X. Gao, Bearing fault diagnosis based on SVD feature extraction and transfer learning classification, in: Proceedings of Prognostics and System Health Management Conference, 2015, pp. 1–6.

[44] A. Gretton, K. Borgwardt, M. Rasch, B. Schlkopf, A. Smola, A kernel two-sample test, Journal of Machine Learning Research 13 (2012) 723–773.

[45] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, B. K. Sriperumbudur, Optimal kernel choice for large-scale two-sample tests, Curran Associates, Inc., 2012.

[46] Y. Li, K. Swersky, R. Zemel, Generative moment matching networks, in: Proceedings of 32nd International Conference on Machine Learning, 2015, pp. 1718–1727.

[47] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of 26th Annual Conference on Neural Information Processing Systems, Vol. 2, 2012, pp. 1097–1105.

[48] Y. Lei, F. Jia, J. Lin, S. Xing, S. X. Ding, An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data, IEEE Transactions on Industrial Electronics 63 (5) (2016) 3137–3147.

[49] B. Liu, J. Liu, X. Bai, H. Lu, Regularized hierarchical feature learning with non-negative sparsity and selectivity for image classification, in: Proceedings of 22nd International Conference on Pattern Recognition, 2014, pp. 4293–4298.

[50] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of 32nd International Conference on Machine Learning, Vol. 1, Lile, France, 2015, pp. 448–456.

[51] G. E. Dahl, T. N. Sainath, G. E. Hinton, Improving deep neural networks for LVCSR using rectified linear units and dropout, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8609–8613.

[52] H. Zhu, T. Rui, X. Wang, Y. Zhou, H. Fang, Fault diagnosis of hydraulic pump based on stacked autoencoders, in: Proceedings of 12th IEEE International Conference on Electronic Measurement & Instruments, 2015, pp. 58–62.

[53] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 .

[54] K. Zhang, B. Schölkopf, K. Muandet, Z. Wang, Domain adaptation under target and conditional shift, in: Proceedings of 30th International Conference on Machine Learning, Vol. 28, 2013, pp. 819–827.

[55] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, Journal of Machine Learning Research 9 (2010) 249–256.

[56] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.

[57] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 .

[58] W. A. Smith, R. B. Randall, Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study, Mechanical Systems and Signal Processing 6465 (2015) 100–131.

[59] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, arXiv preprint arXiv:1702.05464 .

[60] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, T. Zhang, Deep model based domain adaptation for fault diagnosis, IEEE Transactions on Industrial Electronics 64 (3) (2017) 2296–2305.

[61] L. Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2625.

[62] W. Li, S. Zhang, G. He, Semisupervised distance-preserving self-organizing map for machine-defect detection and classification, IEEE Transactions on Instrumentation and Measurement 62 (5) (2013) 869–879.

[63] B. J. van Wyk, M. A. van Wyk, G. Qi, Difference histograms: A new tool for time series analysis applied to bearing fault diagnosis, Pattern Recognition Letters 30 (6) (2009) 595–599.

[64] B. Muruganatham, M. A. Sanjith, B. Krishnakumar, S. A. V. Satya Murty, Roller element bearing fault diagnosis using singular spectrum analysis, Mechanical Systems and Signal Processing 35 (12) (2013) 150–166.

[65] W. Du, J. Tao, Y. Li, C. Liu, Wavelet leaders multifractal features based fault diagnosis of rotating mechanism, Mechanical Systems and Signal Processing 43 (12) (2014) 57–75.

[66] X. Jin, M. Zhao, T. W. S. Chow, M. Pecht, Motor bearing fault diagnosis using trace ratio linear discriminant analysis, IEEE Transactions on Industrial Electronics 61 (5) (2014) 2441–2451.

[67] X. Zhang, Y. Liang, J. Zhou, Y. zang, A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM, Measurement 69 (2015) 164–179.

Table 1
Default parameters of the proposed method and the experimental setting.

Parameter	Value	Parameter	Value
Epoch number	2000	N_{sou}	5/10/20/30/50/100
Learning rate	0.001	N_{tar}	5/10/20/30/50/100
F_N	10	N_{test}	400
F_L	10	N_{input}	500

Table 2
The rolling bearing dataset information.

Class Label	Fault Location	Fault Size (mil)	Load (hp)	Sample Length
1	N/A (H)	0	$0, 1, 2, 3$	N_{input}
2	IF	7	$0, 1, 2, 3$	N_{input}
3	IF	14	$0, 1, 2, 3$	N_{input}
4	IF	21	$0, 1, 2, 3$	N_{input}
5	BF	7	$0, 1, 2, 3$	N_{input}
6	BF	14	$0, 1, 2, 3$	N_{input}
7	BF	21	$0, 1, 2, 3$	N_{input}
8	OF	7	$0, 1, 2, 3$	N_{input}
9	OF	14	$0, 1, 2, 3$	N_{input}
10	OF	21	$0, 1, 2, 3$	N_{input}

Table 3
The information of the transfer tasks in this paper.

Transfer Task	Source Domain	Target Domain	No. of Samples	No. of Samples
	(Load)	(Load)	from Source Domain	from Target Domain
T_{01}	0	1	$10 \times N_{sou}$	$10 \times N_{tar}$
T_{02}	0	2	$10 \times N_{sou}$	$10 \times N_{tar}$
T_{03}	0	3	$10 \times N_{sou}$	$10 \times N_{tar}$
T_{30}	3	0	$10 \times N_{sou}$	$10 \times N_{tar}$
T_{31}	3	1	$10 \times N_{sou}$	$10 \times N_{tar}$
T_{32}	3	2	$10 \times N_{sou}$	$10 \times N_{tar}$

Table 4
Comparisons of classification accuracy of related researches on the same rolling bearing dataset.

Training with	Method	Number of	Testing Accuracy
		Fault Classes	on Target Domain
Labeled Target Domain Samples	[62]	4	95.8%
	[65]	10	88.9%
	[66]	10	92.5%
	[67]	11	97.91%
	[48]	10	99.66%
Labeled Source Domain and Unlabeled Target Domain Samples	[60]	4	94.73%
	Proposed (default)	10	**99.17**%
	Proposed (enhanced)	10	**99.76**%

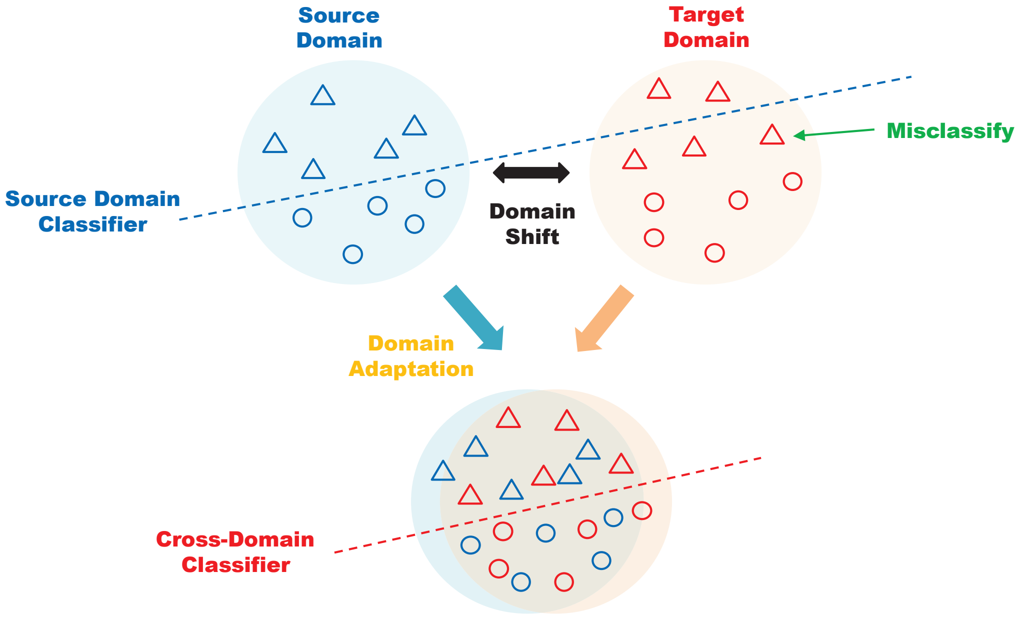

Fig. 1. Illustration for domain adaptation. The triangles and circles denote two different classes.

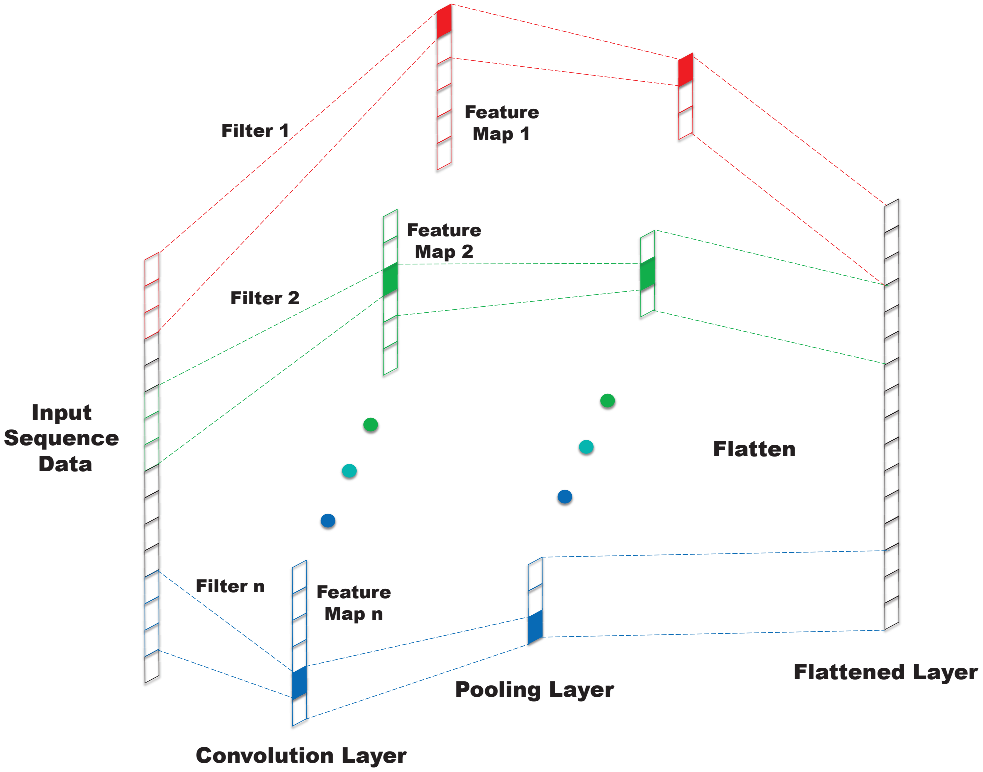

Fig. 2. Illustration for 1D CNN operation.

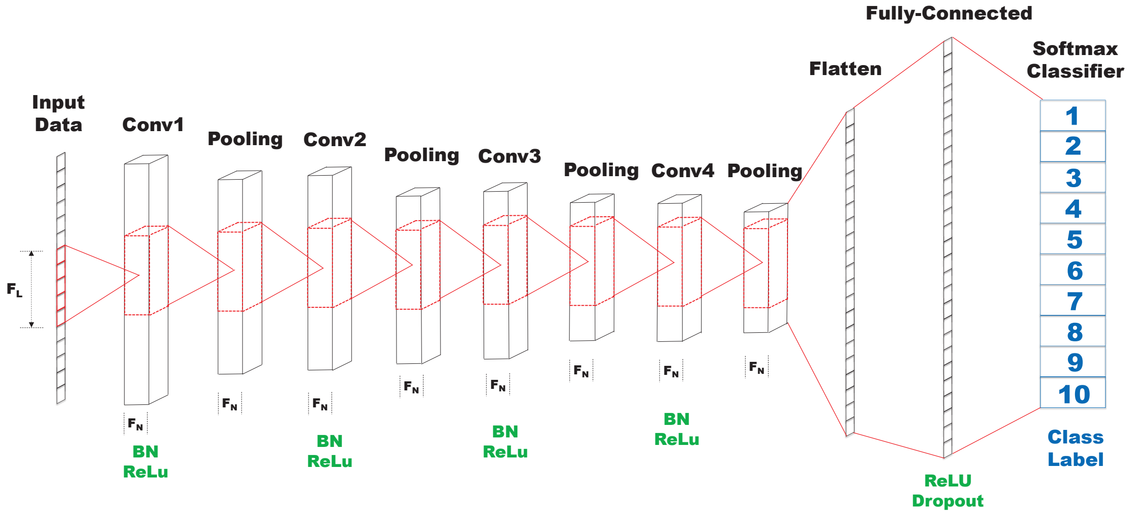

Fig. 3. Proposed deep learning architecture for fault diagnosis. Conv1 to Conv4 denote the 4 convolutional layers, which are followed by batch normalizations (BN) and ReLU activations.

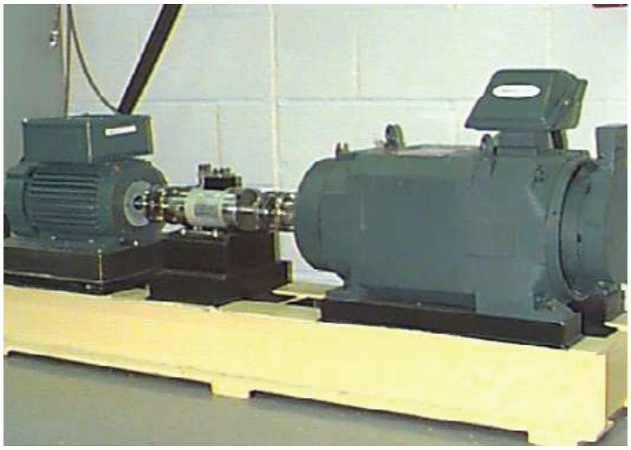

Fig. 4. The bearing test rig used in the experiments [58].

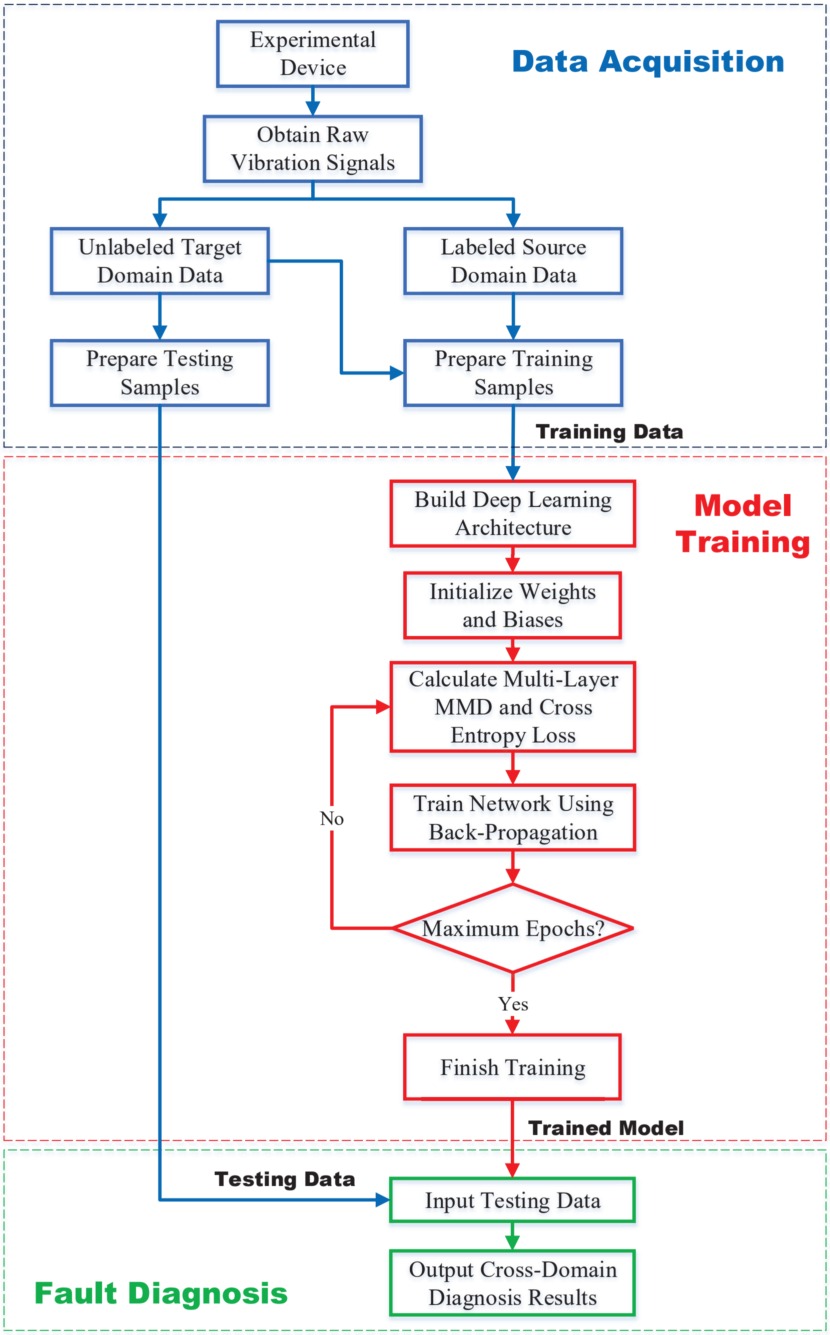
Data Acquisition
Experimental Device
Obtain Raw Vibration Signals
Unlabeled Target Domain Data
Labeled Source Domain Data
Prepare Testing Samples
Prepare Training Samples
Training Data
Model Training
Build Deep Learning Architecture
Initialize Weights and Biases
Calculate Multi-Layer MMD and Cross Entropy Loss
Train Network Using Back-Propagation
No
Maximum Epochs?
Yes
Finish Training
Trained Model
Testing Data
Input Testing Data
Output Cross-Domain Diagnosis Results
Fault Diagnosis

Fig. 5. Flow chart of the proposed cross-domain fault diagnosis method.

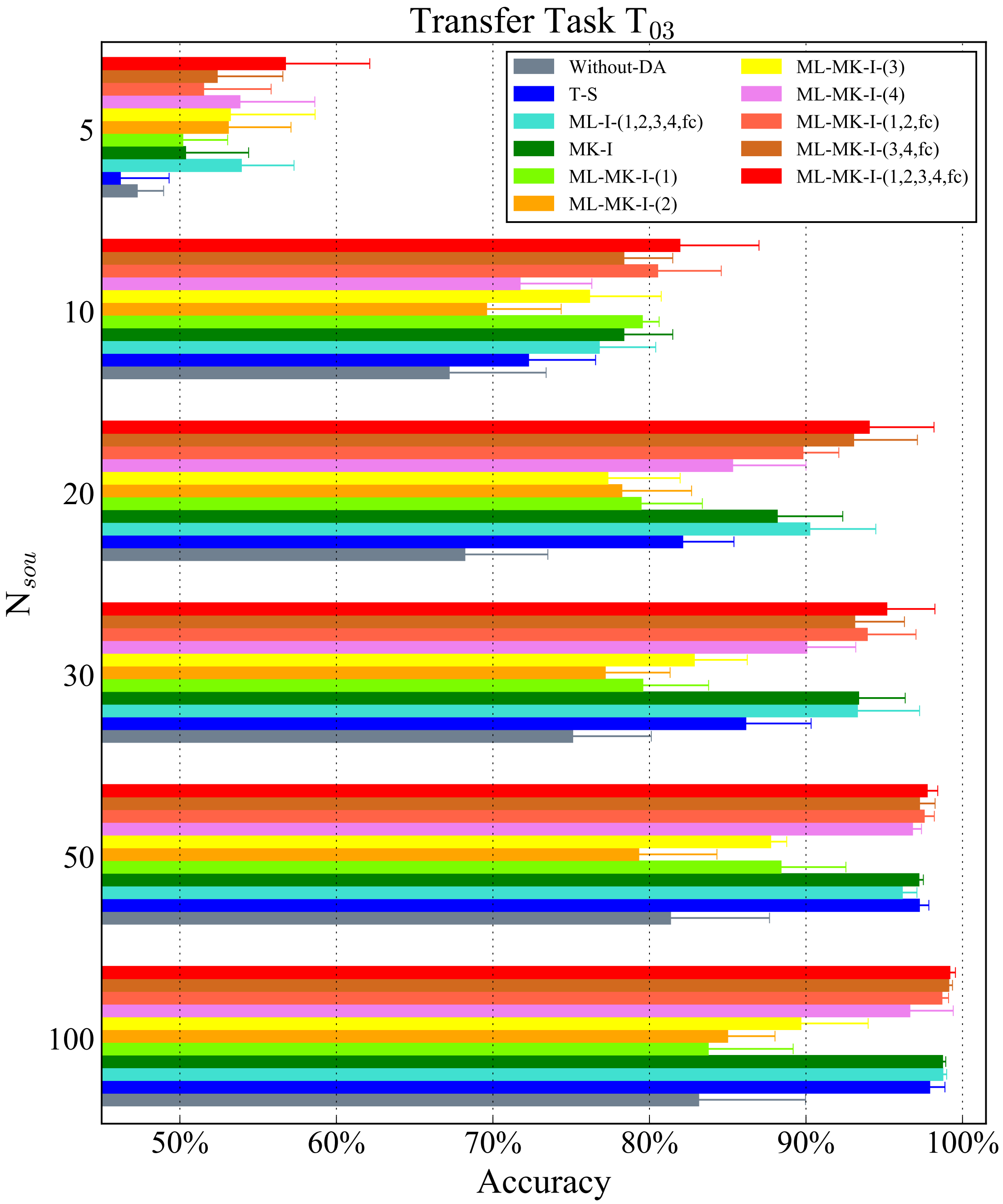

Fig. 6. Testing diagnosis results on the target domain samples in the task T_{03}. Different methods are evaluated using different amount of labeled source domain data.

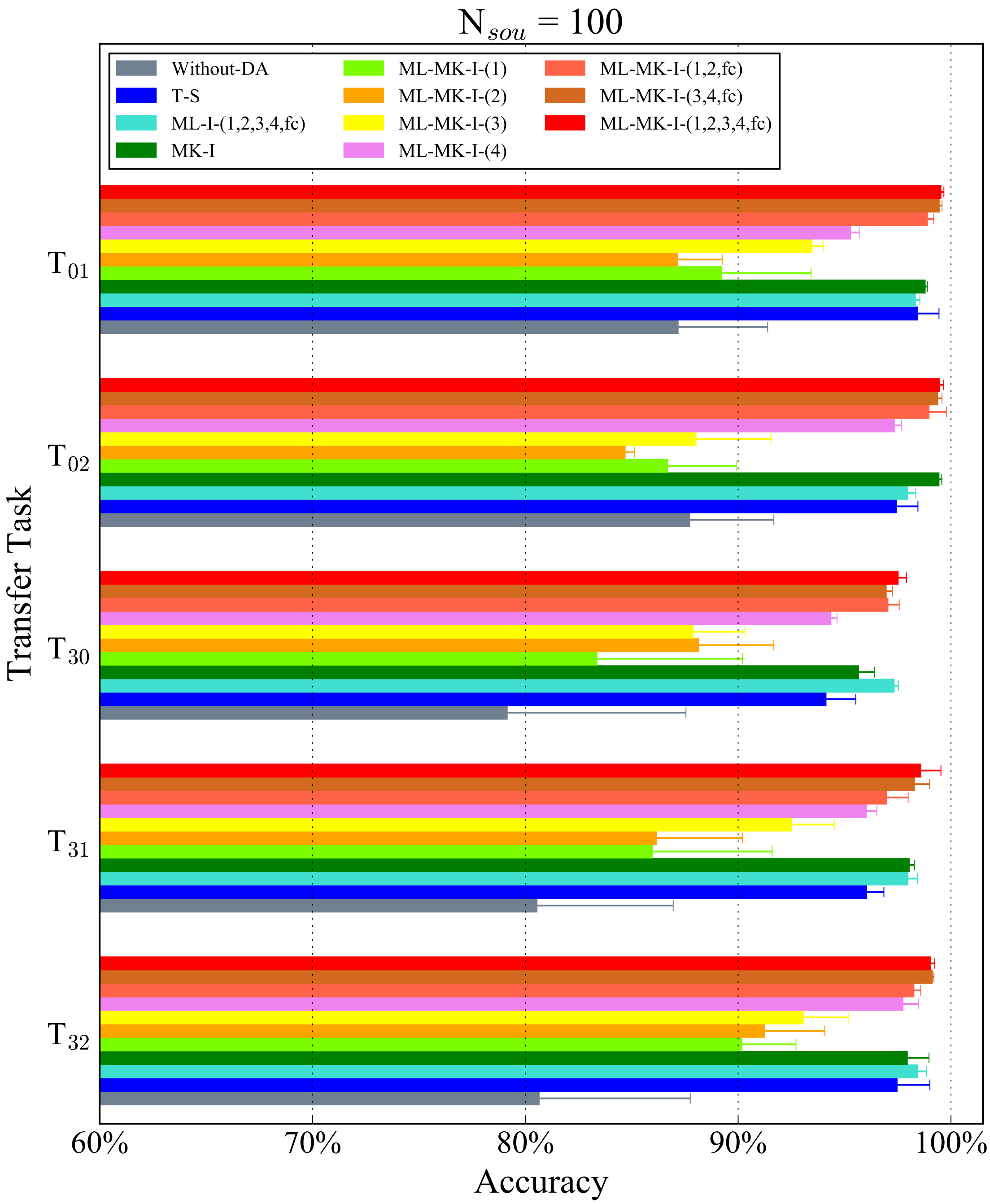

Fig. 7. Testing diagnosis results on the target domain samples in five transfer tasks, i.e. T_{01}, T_{02}, T_{30}, T_{31} and T_{32}. $N_{sou} = 100$ is used.

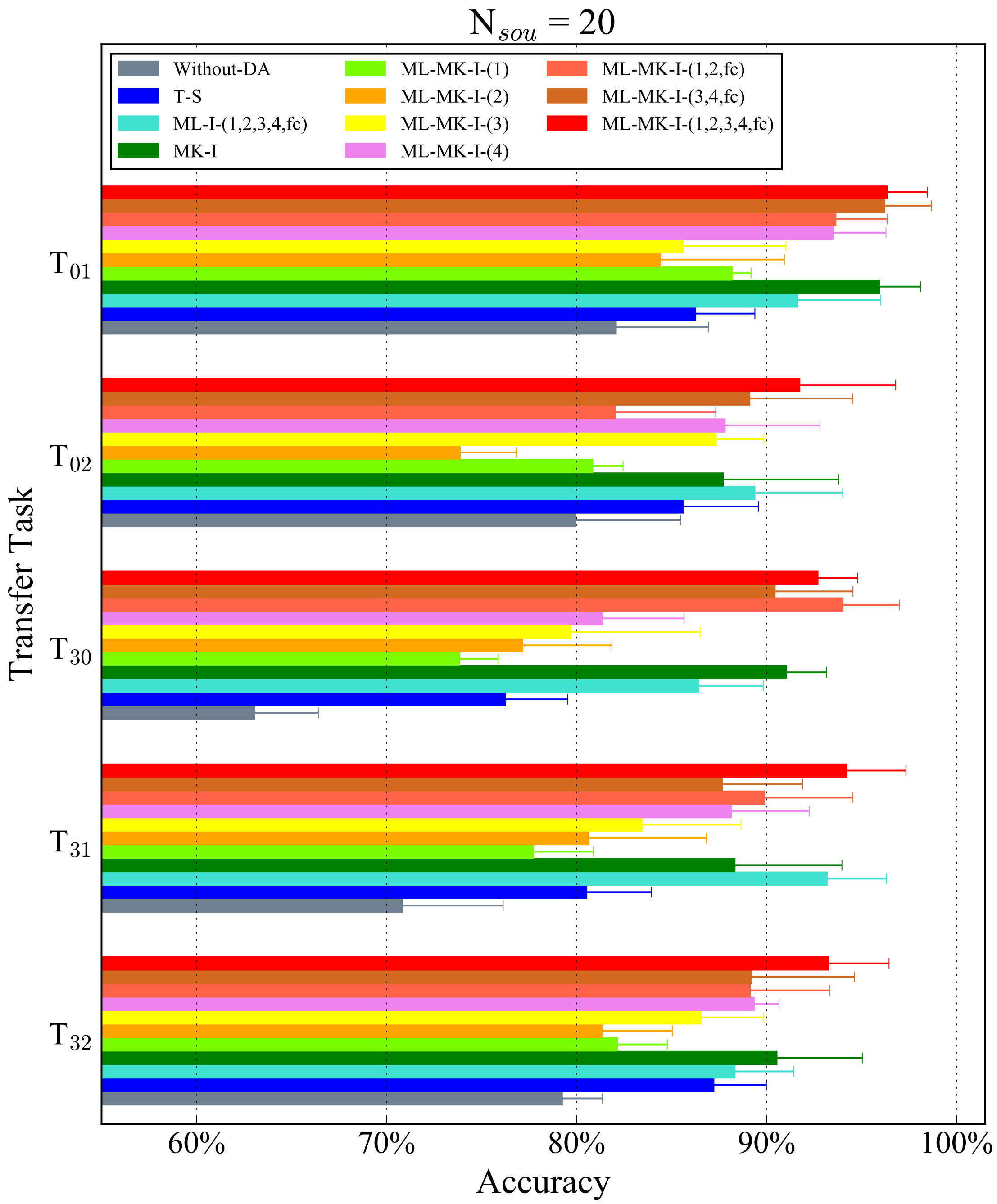

Fig. 8. Testing diagnosis results on the target domain samples in five transfer tasks, i.e. T_{01}, T_{02}, T_{30}, T_{31} and T_{32}. $N_{sou} = 20$ is used.

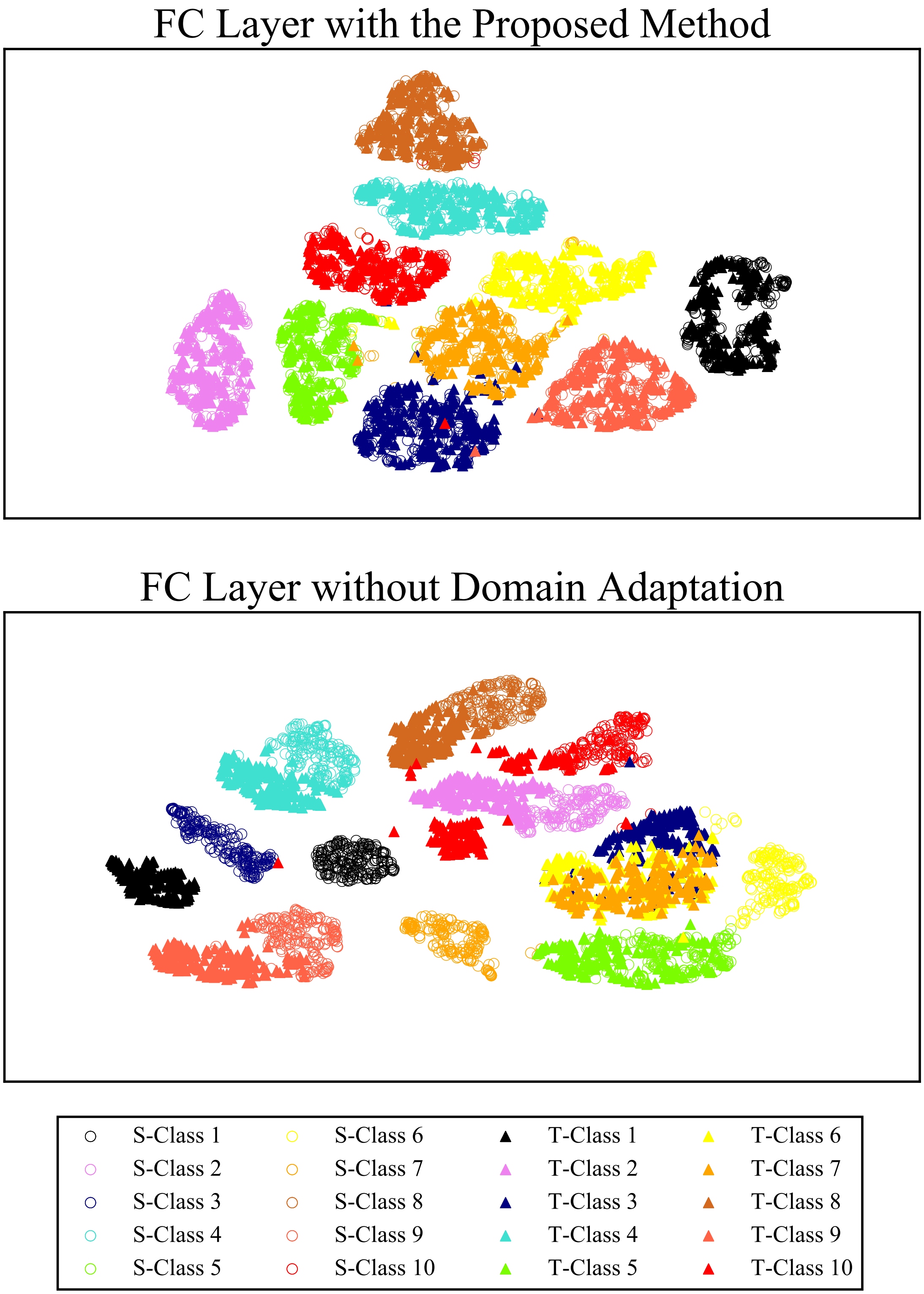

Fig. 9. Visualization of the features in the fully-connected (FC) layer in the T_{03} task. S- and T- in the legend denote the data from the source and target domains respectively.

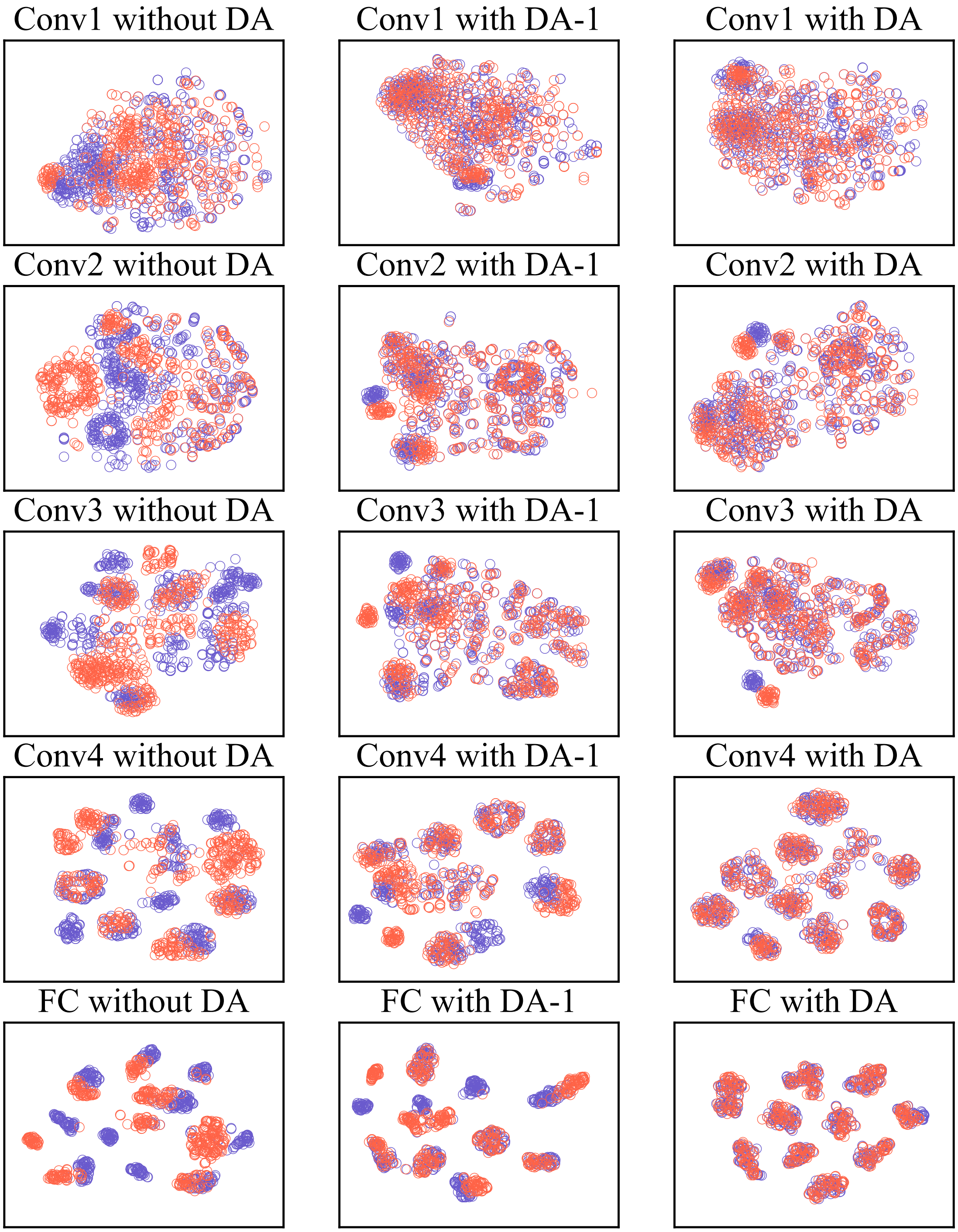

Fig. 10. Visualization of the two domains in each layer of the proposed network in the T_{03} task. Blue circle: Source domain; Red circle: Target domain. Conv1 to Conv4 denote the four convolutional layers respectively. FC represents the fully-connected layer. The three columns show the visualizations using different method. For instance, Conv1 without DA indicates the Without-DA method; Conv1 with DA-1 represents the ML-MK-I-(1) method; Conv1 with DA is by the proposed method ML-MK-I-(1,2,3,4,fc).

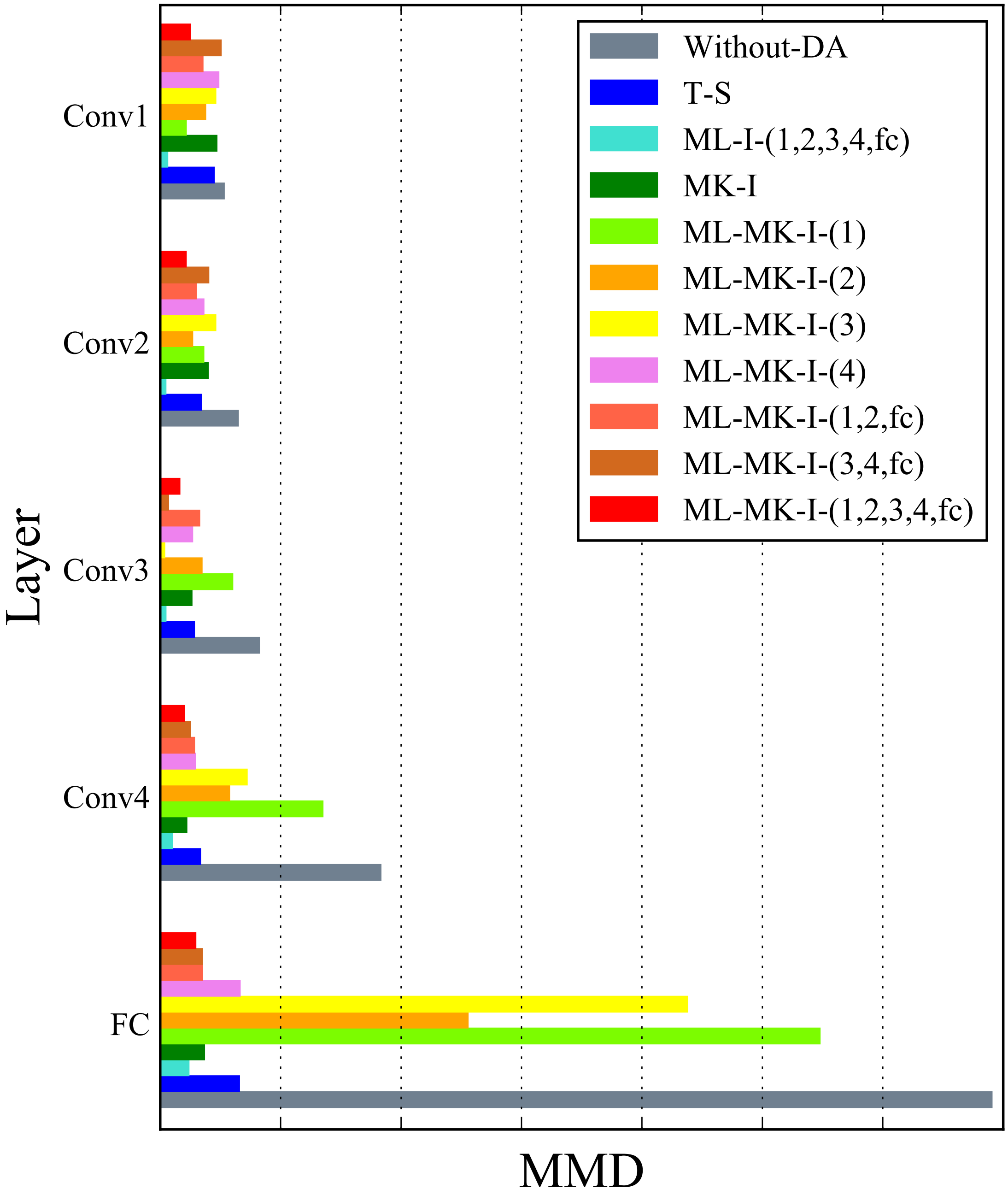

Fig. 11. Distribution MMD in each layer between the source and target domains in task T_{03} using different methods. The abbreviations are the same with those in Figure 10.

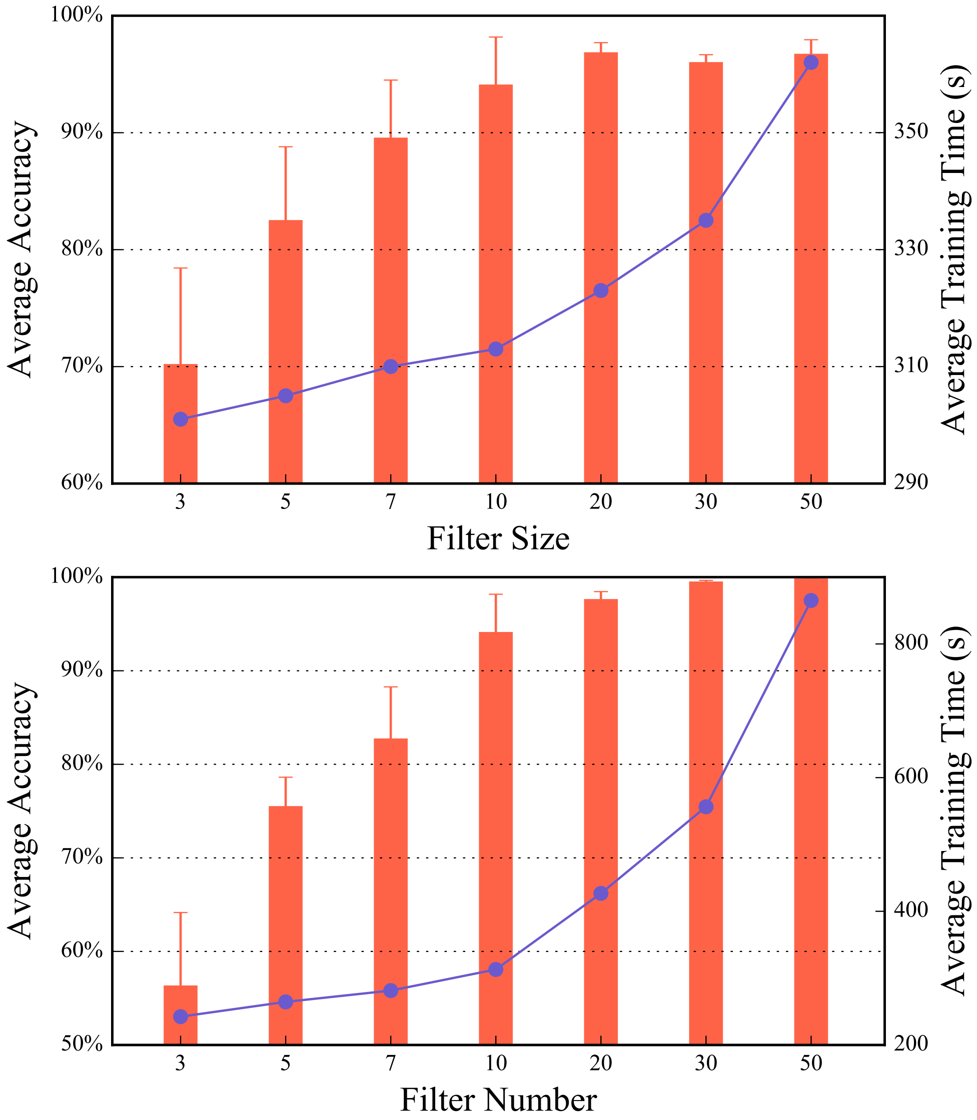

Fig. 12. Effects of the convolutional filter size and number on the testing diagnosis accuracy in task T_{03}. The red bars denote the average testing accuracies, and the blue lines represent the average training time.