Intelligent Fault Identification Based on Multisource Domain Generalization Towards Actual Diagnosis Scenario

Huailiang Zheng , Rixin Wang , Yuantao Yang, Yuqing Li, and Minqiang Xu

Abstract—The data-driven diagnosis methods based on conventional machine-learning techniques have been widely developed in recent years. However, the assumption of conventional methods that the training and test data should be identically distributed is usually unsatisfied in actual diagnosis scenario. While there are several existing works that have been studied to construct diagnosis models by transfer learning methods, most of them are only focused on learning from a single source. Actually, how to discover effective and general diagnosis knowledge from multiple related source domains and further generalize the learned knowledge to new target tasks is crucial to data-driven fault diagnosis. To this end, this paper proposes a novel intelligent fault identification method based on multiple source domains. First, the method describes the discriminant structure of each source domain as a point of Grassmann manifold using local Fisher discriminant analysis. Through preserving the within-class local structure, local Fisher discriminant analysis can learn effective discriminant directions from multimodal fault data. Second, the mean subspace of source domains is computed on the Grassmann manifold through Karcher mean. The mean subspace can be viewed as a representation of the general diagnosis structure that can facilitate the construction of the diagnosis model for the target domain. Experiments on bearing fault diagnosis tasks verify the effectiveness of the proposed method.

Index Terms—Actual diagnosis scenario, data-driven fault diagnosis, multisource domain generalization.

NOMENCLATURE

Notations

x_i	Sample in original space.
y_i	Label of x_i.
D	Dimensionality of original space.
d	Dimensionality of subspace.
C	Number of classes.
k	Number of nearest neighbors.
σ_i	Local scaling of heat kernel.
$\mathbf{T}_{\mathrm{LFDA}}$	Optimal transformation of LFDA.
$\mathbf{T}_{\mathrm{LFDA}}^{\perp}$	Orthogonal discriminant vectors of LFDA.
$\mathcal{G}(d, \mathrm{D})$	Grassmann manifold.
$d_{\mathcal{G}}(S, T)$	Geodesic distance between point S and T on Grassmann manifold.
θ_i	Principle angles of two subspaces.
m	Total number of source domains.
$\mathcal{D}_s^l$	lth source domain/dataset.
$\mathcal{D}_t$	Target domain/dataset.

I. INTRODUCTION

FAULT diagnosis is crucial to guarantee reliability and avoid significant economic loss in industry application. In the last years, data-driven intelligent fault diagnosis methods have been widely developed and applied, due to the development of advanced sensing technologies and its advantage that need not construct complicated physical model over model-based methods [1]. Recent research works about data-driven intelligent diagnosis methods mainly focus on two aspects. The first aspect is to propose new feature extraction methods based on advanced signal processing techniques, such as spectral kurtosis method [2], entropy-based feature [3], and modified empirical mode decomposition method [4]. The second aspect is to utilize and modify different machine-learning methods to learn the mapping relationship between features and fault modes automatically, such as support vector machine (SVM) [5], k nearest neighbor (KNN) [6], and neural network [7]. Recently, deep learning based methods are widely studied and applied to addressing fault diagnosis problem, which can learn hierarchical representations from raw input without using hand-crafted features [8]–[16]. For example, Lei *et al.* [8] have used sparse filtering to directly learn features from bearing vibration signals; then, the Softmax regression is employed to classify the health conditions. Yuan *et al.* [12] proposed a novel variable-wise weighted stacked auto-encoder, which can give better diagnosis performance than the traditional multilayer neural networks and SAE on an industrial debutanizer column process. Li *et al.* [11] proposed a fault diagnosis and isolation method for wind turbines, which implements long short-term memory networks for residual generator and applies the random forest (RF) algorithm for decision making.

However, these conventional supervised learning methods [5]–[7] follow a common assumption that training and test data must be drawn from the same distribution. If this assumption does not hold, the generalization ability of these methods will drop dramatically. Although deep learning based diagnosis methods can automatically learn remarkable features from raw signal, the generalization ability to congeneric diagnosis problems of the features learned by deep models is not discussed and guaranteed. Meanwhile, the identically distributed assumption should also be satisfied in the supervised identification stage because the conventional supervised methods are usually applied in the last layer of these deep models, such as neural network [10], Softmax regression [8], [9], and RF [11]. Unfortunately, for actual diagnosis problem, holding this assumption is very difficult because it means that a training dataset should be collected before diagnosing from the same machine under the same operating condition and even noise environment. Therefore, those intelligent fault diagnosis methods cannot construct effective fault identification models for target machine in actual diagnosis scenario due to the unavailable of training data with same distribution. Moreover, an ideal and significant diagnosis model should not only accurately classify fault modes in a specific dataset, but also show a remarkable generalization ability in related diagnosis tasks.

Actually, the fault data for training identification model are usually collected from different operating conditions or even other same-type machines. Although the *a priori* distribution estimation of test data is unknown and the distribution of test data might be different from training data, they are similar and related to each other because of the same working principle and the similar failure mechanism. Based on the related but distribution-different data, several research works have been studied to build fault diagnosis models using transfer learning or domain adaptation methods [17]–[22]. Shen *et al.* [21] have applied TrAdaBoost, a representative instance-based transfer learning method, to improve diagnosis performance of rolling element bearing under different operation conditions. Similarly, Xie *et al.* [20] proposed to enhance gearbox diagnosis performance by reusing historical data under various operating conditions, and a feature-based transfer method—transfer component analysis—was utilized to transfer knowledge from different operating conditions. Following the feature-based transfer strategy, several research works combined deep learning and transfer learning for machinery fault diagnosis. Lu *et al.* [19] proposed a deep domain adaptation method under the auto-encoder framework and applied this method to bearing and gearbox fault diagnosis problem. A similar method based on sparse auto-encoder is also proposed by Wen *et al.* [17]. Zhang *et al.* [18] presented a transfer learning approach for improving bearing fault diagnosis performance, which can transfer partial parameters from the neural networks trained using enough source data to the neural networks of the target task. Zhang *et al.* [22] designed a convolutional neural network based diagnosis algorithm that achieves high identification accuracy for bearing fault diagnosis under noisy environment and different working loads. The basic idea of these research works is to learn and transfer the diagnosis knowledge from training data to test data and improve the generalization ability of the identification model on target dataset further. In contrast to traditional machine-learning techniques that try to learn a new task from scratch, transfer learning or domain adaptation borrows knowledge from source domains to facilitate learning in a target domain, which has shown promising results in cross-domain learning applications [23], [24].

The abovementioned diagnosis methods using transfer learning have proved the feasibility of promoting diagnosis performance by the manner of knowledge transfer [17]–[22]. But they can only consider one single source in learning process and may suffer some deficiencies in actual diagnosis scenario. First, the premise of effective transfer is that the source and target domains should be directly related [24]. However, how to ensure positive transfer has not been discussed for fault diagnosis problem. One strategy to decrease the risk for negative transfer is to borrow knowledge from multiple related sources [25], [26]. By this way, the chance to learn beneficial diagnosis knowledge closely related to the target machine significantly increases. Second, in actual diagnosis scenario, the samples of target machine under each fault category cannot be collected before diagnosing, and usually only normal samples are available for training diagnosis model. Hence, the distribution discrepancies between the target domain and source domains cannot be estimated for choosing the most related source using current multisource transfer methods [25], [27]–[31]. The fault diagnosis based on multiple sources is more prone to a domain generalization problem than a domain adaptation problem. Therefore, how to discover general diagnosis knowledge from multiple sources [32], [33] and apply these knowledge to a new diagnosis task is the crucial issue for data-driven fault diagnosis.

In light of the above discussion, this paper proposes a novel intelligent fault identification method based on multiple source domains. Our proposed method describes the discriminant structure of each source domain that is favorable for fault identification as a point on Grassmann manifold using a dimensionality reduction method, *local Fisher discriminant analysis* (LFDA). Then, the method tries to discover the general diagnosis knowledge from those source domains by constructing the mean subspace on that Grassmann manifold. To our best knowledge, this paper is the first attempt to consider knowledge transfer from multiple sources for machinery fault diagnosis. In contrast to current multisource transfer methods that try to adapt source domains to target domain, our method is more prone to discover and transfer general diagnosis knowledge (common discriminant structure in our method) from source domains to target domain, which is more suitable to actual diagnosis scenario. In addition, two subspace-based domain adaptation methods are close to ours [34], [35]. They also explored the effective manners of knowledge transfer in multiple source scenarios. The first difference of these works from ours is that we use LFDA instead of principal component analysis (PCA) to discover discriminant diagnosis structure of each source. Compared with PCA that identifies the directions of variances maximization without considering label information, LFDA can learn the optimal subspace for discriminant analysis with preserving local structure of the data. By this way, the diagnosis structure that

is associated with discriminant information will be represented more effectively. The second difference is that we directly embed target samples (test samples) to common diagnosis subspace without sampling finite [35] or infinite subspaces [34] along the geodesic flow connected source and target datasets. The main reason is that the target domain cannot be embedded on the Grassmann manifold correctly without the fault data of target domain in training stage. The main contributions of our paper can be summarized as follows.

1) This paper proposes a new data-driven fault identification method that can borrow diagnosis knowledge from multiple related source domains. The data of each source domain can be collected from different operating conditions or other same-type machines. The proposed method is very promising for handling actual fault diagnosis problem than conventional data-driven methods. The manner of knowledge transfer is a new attempt to achieve higher level of intelligent fault identification.

2) The well-designed method in our paper is suitable to address the mechanical fault diagnosis problem. In the proposed method, LFDA is used to learn the optimal discriminant structure of each source domain from multimodal fault data. In addition, because of lacking fault data of the target domain in the training stage, we directly map target samples to common subspace without embedding it on Grassmann manifold.

3) The proposed method can discover general diagnosis knowledge from multiple source domains and apply the knowledge to facilitate new tasks in target domain. The risk of negative transfer in learning process is reduced.

This paper is organized as follows. Preliminaries about our proposed method and problem formulation are described in Section II. Then, Section III presents the proposed method in detail. Experimental results and analysis are illustrated in Section IV. Section V concludes this paper.

II. PRELIMINARIES

In this section, we provide the background theory for the techniques involved in our proposed method.

A. Local Fisher Discriminant Analysis

LFDA is a supervised dimensionality reduction method. Compared with the traditional *Fisher discriminant analysis* (FDA) [36], LFDA can preserve the local structure of the data through combining the idea of *locality preserving projection* [37]. Let $\{(\boldsymbol{x}_i, y_i)\}|_{i=1}^{n}$ be labeled samples with $\boldsymbol{x}_i \in \mathbb{R}^{D}$ and $y_i \in \{1, 2, \ldots, C\}$, where n is the number of samples and C is the number of classes. Let n_c be the number of samples in class c, then $\sum_{c=1}^{C} n_c = n$. The objective of LFDA is to learn a $D \times d$ transformation matrix $\mathbf{T}$ that maps the samples to d-dimensional subspace, where d is the dimension of the embedding space [37].

The form of the objective function of LFDA is the same with FDA, which is defined as follows:

$$\mathbf{T}_{\text{LFDA}} \equiv \underset{\mathbf{T} \in \mathbb{R}^{D \times d}}{\operatorname{argmax}} \left[\operatorname{tr} \left(\left(\mathbf{T}^{\mathrm{T}} S^{(w)} \mathbf{T} \right)^{-1} \mathbf{T}^{\mathrm{T}} S^{(b)} \mathbf{T} \right) \right] \quad (1)$$

where $S^{(w)}$ is the *local* within-class scatter matrix and $S^{(b)}$ is the *local* between-class scatter matrix. They are defined as follows respectively:

$$S^{(w)} \equiv \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(w)} \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right) \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right)^{\mathrm{T}} \quad (2)$$

$$S^{(b)} \equiv \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(b)} \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right) \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right)^{\mathrm{T}} \quad (3)$$

where

$$W_{i,j}^{(w)} \equiv \begin{cases} A_{i,j}/n_c, & \text{if } y_i = y_j = c \\ 0, & \text{if } y_i \neq y_j \end{cases} \quad (4)$$

$$W_{i,j}^{(b)} \equiv \begin{cases} A_{i,j} \left(1/n - 1/n_c \right), & \text{if } y_i = y_j = c \\ 1/n, & \text{if } y_i \neq y_j \end{cases}. \quad (5)$$

$\mathbf{A}$ is the similarity matrix with the element $A_{i,j}$ having the weight of the edge jointing $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. In LFDA, the similarity matrix $\mathbf{A}$ is constructed in a supervised manner. A standard choice of $\mathbf{A}$ is heat kernel with local scaling

$$A_{i,j} =$$
$$\begin{cases} \exp \left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma_i \sigma_j} \right), & \text{if } \boldsymbol{x}_i \in \mathcal{N}_k^s \left(\boldsymbol{x}_j \right) \vee \boldsymbol{x}_j \in \mathcal{N}_k^s \left(\boldsymbol{x}_i \right) \\ 0, & \text{otherwise} \end{cases}$$
$$(6)$$

where $\mathcal{N}_k^s(\boldsymbol{x}_j)$ is the set of KNN of $\boldsymbol{x}_j$ searched from the samples of same class. σ_i represents the local scaling of the data samples around $\boldsymbol{x}_i$, which is determined by $\sigma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i^{(k)}\|$, where $\boldsymbol{x}_i^{(k)}$ is the kth nearest neighbor of $\boldsymbol{x}_i$.

The optimal transformation $\mathbf{T}_{\text{LFDA}}$ is computed by solving a generalized eigenvalue problem of $S^{(b)}$ and $S^{(w)}$. The detailed implementation and code can be found in the literature [37]. The orthogonal discriminant vectors $\mathbf{T}_{\text{LFDA}}^{\perp}$ of LFDA are given by the QR decomposition of $\mathbf{T}_{\text{LFDA}}$, that is $\mathbf{T}_{\text{LFDA}} = \mathbf{T}_{\text{LFDA}}^{\perp} R$ [38]. The columns of $\mathbf{T}_{\text{LFDA}}^{\perp}$ are orthogonal to each other, and $\mathbf{T}_{\text{LFDA}}^{\perp}$ can be treated as a point on a Grassmann manifold. LFDA is a localized extension to FDA. Through preserving within-class local structure, LFDA do not restrict the class-conditional distributions to be Gaussian and can reduce the dimensionality of multimodal data appropriately. In contrast to linear unsupervised dimensionality reduction methods, such as PCA, LFDA is more prone to find the projection with maximum discrimination and is an appropriate choice for handling classification problems of multimodal data.

B. Grassmann Manifold

A Grassmann manifold, denoted by $\mathcal{G}(d, D)$, consists of all d-dimensional subspaces embedded in D-dimensional Euclidean space $\mathbb{R}^{D} (0 < d < D)$, where each subspace maps to a unique point on the manifold [39], [40]. Each point in $\mathcal{G}(d, D)$ can be represented using a $D \times d$ orthonormal matrix, which forms a basis for that subspace. Given two points $\boldsymbol{S}$ and $\boldsymbol{T}$ on $\mathcal{G}(d, D)$, as shown in Fig. 1, the geodesic distance can be computed by
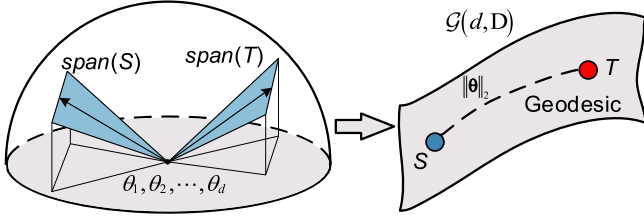
Fig. 1. Structure of Grassmann manifold $\mathcal{G}(d, \mathrm{D})$ and geodesic distance between two points on $\mathcal{G}(d, \mathrm{D})$.
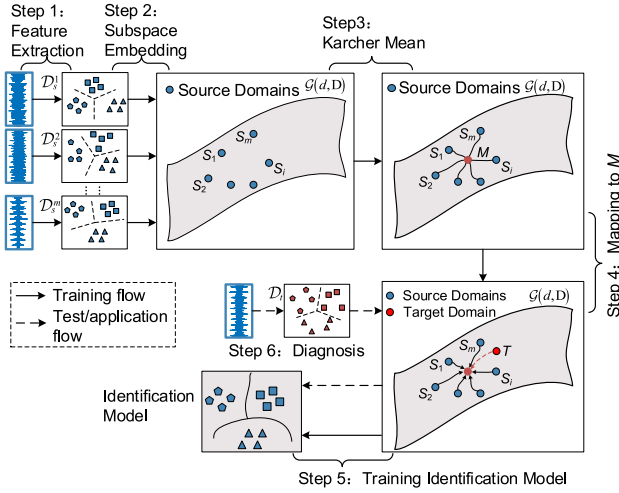

Fig. 2. Framework of the proposed diagnosis method.

$d_{\mathcal{G}}(\boldsymbol{S}, \boldsymbol{T}) = \|\boldsymbol{\theta}\|_2 = \sum_i (\theta_i^2)^{1/2}$, where $\theta_i \in [0, \pi/2]$ denotes the principle angles of the two subspaces. The principle angle vector $\boldsymbol{\theta}$ is computed by decomposing $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{T}$ using its SVD.

C. Problem Formulation

Let $\mathcal{D}_s^1, \mathcal{D}_s^2, \ldots, \mathcal{D}_s^m$ be m ($m > 1$) source domains and $\mathcal{D}_t$ be the target domain. The dataset from the lth source domain, denoted by $\mathcal{D}_s^l = \{(\boldsymbol{x}_i^l, y_i^l)\}|_{i=1}^{n_l}$, is fully labeled, where $\boldsymbol{x}_i^l \in \mathbb{R}^{\mathrm{D}}$ and $y_i^l \in \{1, 2, \ldots, C\}$. D is the dimensionality of feature vector, and C is the number of health conditions. In the diagnosis scenario of this paper, each domain has the same dimensionality D and condition number C. In training stage, some labeled normal samples of target domain $\mathcal{D}_t = \{(\boldsymbol{x}_i^t, y_i^t)\}|_{i=1}^{n_t}$ are available as well. Our objective is to identify the samples of target domain by learning diagnosis knowledge from $\mathcal{D}_s^1, \mathcal{D}_s^2, \ldots, \mathcal{D}_s^m$.

It is worth to mention that we assume the distribution of target domain is different from that of each source domain. In fault diagnosis field, this is a reasonable assumption. Actually, when we aim to identify the fault modes of a machine, the training data available are usually collected from different operating conditions or other same-type machines.

III. PROPOSED INTELLIGENT FAULT IDENTIFICATION METHOD

A. General Procedure of the Proposed Method

The framework of this proposed method is shown in Fig. 2 and the general procedures are summarized as follows.

Step 1: Extract effective fault features from monitoring signals to describe the health condition of the machine. After feature extraction, a data matrix denoted by $[\boldsymbol{x}_1^l | \boldsymbol{x}_2^l |, \ldots, | \boldsymbol{x}_{n_l}^l]_{\mathrm{D} \times n_l}$ is obtained for each domain.

Step 2: Embed each source domain to a d-dimensional subspace using LFDA. Then, each transformation matrix $\boldsymbol{S}_l$ ($l = 1, 2, \ldots, m$) is viewed as a point of the Grassmann manifold $\mathcal{G}(d, \mathrm{D})$.

Step 3: Compute the mean subspace $\boldsymbol{M} \in \mathbb{R}^{\mathrm{D} \times d}$ of source domains on the Grassmann manifold.

Step 4: Map the samples of each domain to mean subspace $\boldsymbol{M}$ by $\boldsymbol{M}^{\mathrm{T}}[\boldsymbol{x}_1^l | \boldsymbol{x}_2^l |, \ldots, | \boldsymbol{x}_{n_l}^l]$ where $l = 1, 2, \ldots, m$. Note that the normal data of target domain are also mapped to $\boldsymbol{M}$ in this step.

Step 5: The training dataset (the samples of m source domains and normal samples of target domain) are used to train a fault identification model for target domain using a classification algorithm. SVM is applied in the following experiment, but it is not confined to it.

Step 6: At last, the unlabeled samples of target domain are classified using the fault diagnosis model trained in *Step 5*. Note that the original monitoring signals collected from target machine are processed by feature extraction step (*Step 1*) and mapping step (*Step 4*) before feeding into the identification model.

B. Discovering Discriminant Structure Using LFDA

Actually, the essential problem of fault identification is to discover the effective diagnosis knowledge from massive training data. For a pattern recognition problem, the diagnosis knowledge is directly related to discriminant structure of each category in the data. Moreover, the discriminant structure can be represented by optimal discriminant directions. In our method, LFDA is applied to learn orthonormal discriminant directions from each domain, and there are two main reasons.

First, the discriminant directions are learned by a dimensionality reduction method, hence different methods will influence the final diagnosis results. We hope that the learning method can utilize as much information as possible. In the diagnosis scenario of this paper, each source domain has label information that is possible of achievement in engineering. For example, the training data are sampled from fault simulation experiments in the laboratory or accumulated from labeled monitoring data of industrial process. Thus, a supervised dimensionality reduction method is selected to obtain more discriminating directions for fault classification by considering the label information of each source domain. Second, the distribution of each domain is empirically estimated by a dataset in diagnosis problem. However, in order to describe fault characteristics fully, the samples may be collected from different degradation stages of a machine. Two two-dimensional (2-D) examples of rolling element bearing dataset are shown in Fig. 3. It can be observed that the samples in a fault mode may form several separate clusters, and the distribution of the dataset shows multimodal and non-Gaussian property. LFDA just fit this dimensionality reduction scenario and could discover more effective discriminant
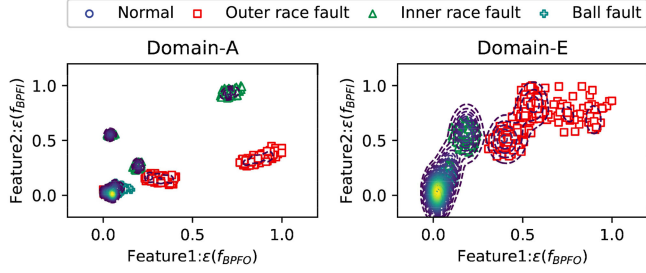
Fig. 3. Two examples of multimodal fault data (the information about the dataset is described in Section IV-A).

Algorithm 1: Mean Subspace Learning of Multisource Domain Generalization for Fault Identification.

Input: m source domains $\mathcal{D}_s^1, \mathcal{D}_s^2, \ldots, \mathcal{D}_s^m$; target domain $\mathcal{D}_t$; subspace dimension d; number of nearest neighbor k.

Output: mean subspace M; mapped training samples in M.

1: **For** $l = 1:m$
    Compute $\mathbf{T}_{\mathrm{LFDA}}^l$ of $\mathcal{D}_i^l$ by optimizing (1)
    Compute S_l by QR decomposition of $\mathbf{T}_{\mathrm{LFDA}}^l$
  End

2: Compute the mean subspace M by optimizing (7)

3: **For** $l = 1:m$
    Map samples of l-th domain to M space by $M^{\mathrm{T}}\left[x_1^l\big|x_2^l\big|,\ldots,\big|x_{n_l}^l\right]$
  End

4: Map samples of target domain to M space by $M^{\mathrm{T}}\left[x_1^t\big|x_2^t\big|,\ldots,\big|x_{n_t}^t\right]$

---

directions than other supervised dimensionality reduction method, such as *orthogonal linear discriminant analysis* (OLDA) [38].

It should be noted that it is not the only way to describe discriminant knowledge using subspace basis in multisource transfer methods. Using discriminant model in decision level has also been applied in some research works [41], [42]. However, one advantage of the subspace-based method is that the process of finding common characteristics can be transformed to an optimization problem on a Grassmann manifold, which has superior behaviors in finding optimal subspace than other subspace-based algorithms [39].

C. Computing Mean Subspace

By subspace embedding, each domain $\mathcal{D}_s^l$ is treated as a point $S_l \in \mathbb{R}^{\mathrm{D} \times d}$ on a Grassmann manifold $\mathcal{G}(d, \mathrm{D})$. S_l $(l = 1, 2, \ldots, m)$, the $\mathrm{D} \times d$ column orthonormal matrices, represent the optimal fault discriminant directions of m source domains, respectively. Note that all tasks corresponding to the m sources are the same diagnosis problem, and thus, there should be general diagnosis knowledge contained in these m sources. More importantly, the general diagnosis knowledge may be helpful to a new and related target task. Therefore, in actual diagnosis scenario that fault data of target machine are unavailable, it is vital to discover those potential general diagnosis knowledge from multiple related source domains. To this end, we compute the mean subspace of source domains on the Grassmann manifold, which is denoted as M. The mean subspace M is defined by the following optimization problem that is called Karcher Mean in the literature [43]:

$$\mathrm{argmin}_{M \in \mathbb{R}^{\mathrm{D} \times d}} \quad \sum_{i=1}^{m} d_{\mathcal{G}}^2\left(M, S_i\right)$$

$$\text{s.t.} \quad M^T M = I_d \tag{7}$$

where $d_{\mathcal{G}}(M, S_i)$ is the geodesic distance between M and the ith source domain S_i on Grassmann manifold. It can be computed by the principle angles of the two subspaces, as depicted in preliminaries. This optimization problem on Grassmann is solved by Riemann–Newton method iteratively [43]. Given an initial point M_0, the update formula in iteration is

$$M_{i+1} = \exp_{M_i}\left(\varepsilon \bar{\xi}\right) \tag{8}$$

where $\bar{\xi} = \frac{1}{m}\sum_{j=1}^{m}\exp_{M_i}^{-1}(S_j)$ is the average tangent vector, and $\varepsilon > 0$ is small step size. In the following experiments,

we use the MATLAB toolbox, Manopt [44], to implement this optimization problem.

The mean subspace M is the mean discriminant directions of the m source domains and can be viewed as a representation of the common diagnosis structure. More specifically, the geodesic distance $d_{\mathcal{G}}(M, S_i)$ describes the similarity of discriminant structure between M subspace and S_i subspace. If M is more close to S_i on the manifold $\mathcal{G}(d, \mathrm{D})$, the distance between them will be smaller, and the discriminant structure of M will be more similar to S_i's. Through considering all m sources, Karcher Mean finds the optimal subspace M that minimizes the sum of geodesic distances, and it means that M is the most similar subspace to all source domains. By this way, the problem of finding general diagnosis knowledge has been transformed to an optimization problem on Grassmann manifold. Optimization on Grassmann manifold has proven effective at avoiding bad local minima. More precisely, manifold optimization methods often have better convergence behavior than iterative projection methods, which can be crucial with a nonlinear objective function [40].

After obtaining Karcher Mean M, all training samples are mapped to M subspace, and then applied to train an identification model using a classification algorithm. In the last stage, the identification model is applied on the unseen samples of target domain to identify the health conditions of machinery.

D. Computational Complexity

The mean subspace learning algorithm in our intelligent fault identification method is summarized in Algorithm 1. The computational complexity of the algorithm consists of four parts.

1) Computing $\mathbf{T}_{\mathrm{LFDA}}^l$ s for all m source domains requires $O((\mathrm{D}^2 + \mathrm{D} + k)\sum_{l=1}^{m} n_l^2 + m(\mathrm{D}^3 + d\mathrm{D}^2))$.

2) Computing S_l $(l = 1, 2, \ldots, m)$ by QR decomposition requires $O(m(2\mathrm{D}d^2 - \frac{2}{3}d^3))$.

3) For computing the mean subspace M, the algorithm requires $O((m+1)\mathrm{D}d^2 N_k)$ to solve the optimization problem (7). N_k is the number of iterations that depends on the m points S_ls on Grassmann manifold and step size ε.

4) Mapping the samples of source and target domains needs $O(d\mathrm{D}(\sum_{l=1}^{m} n_l + n_t))$.

In summary, the approximate computational complexity of our subspace learning algorithm is $O((\mathrm{D}^2 + \mathrm{D} + k)N_s^2 +$

Bearing	Pitch diameter	Ball diameter	Contact angle	Number of balls
SKF 6205	39.04 mm	7.94 mm	0°	9
SKF 6203	28.50 mm	6.75 mm	0°	8

TABLE II
EIGHT DIAGNOSIS TASKS

Tasks	Sources	Target	Tasks	Sources	Target
Multi-A	E/F/G/H	A	Multi-E	A/B/C/D	E
Multi-B	E/F/G/H	B	Multi-F	A/B/C/D	F
Multi-C	E/F/G/H	C	Multi-G	A/B/C/D	G
Multi-D	E/F/G/H	D	Multi-H	A/B/C/D	H

$(m + 1)\mathrm{D}d^2N_k + m(\mathrm{D}^3 + d\mathrm{D}^2 + 2\mathrm{D}d^2 - \frac{2}{3}d^3) + d\mathrm{D}n)$,
where $N_s^2 = \sum_{l=1}^{m} n_l^2$, $n = \sum_{l=1}^{m} n_l + n_t$. In general, n_l is much greater than D and d. Therefore, the main part of algorithmic complexity is $O((\mathrm{D}^2 + \mathrm{D} + k)N_s^2)$ that is determined by the number of sources and the number of samples in each domain.

IV. EXPERIMENTAL EVALUATION

A. Data Preparation

To verify the effectiveness, we empirically evaluate the performance of the proposed method on rolling element bearing dataset [45]. The dataset was acquired from bearing data center of Case Western Reserve University. Vibration signals were collected using accelerometers, which were placed at both the drive end and fan end of the motor housing. The sampling frequency of the signal is 12 kHz. In order to simulate a multisource transfer scenario, we set the samples collected from one operation condition as a domain. Thus, four domains, denoted as A/B/C/D respectively, are organized using the experimental data of drive end bearing (SKF 6205-2RS) defects corresponding to 0–3 hp four motor loads. Similarly, other four domains, E/F/G/H, are organized using fan end bearing (SKF 6203-2RS) experimental data. The parameters of two experimental bearings are list in Table I. There are four health conditions of the bearing in each domain: Normal (N), outer race fault (OR), inner race fault (IR), and ball fault (Ball). For each fault condition, vibration signals of three severity levels (0.18, 0.36, and 0.53 mm) are included and viewed as the same class. We sliced the original signals of each domain with overlap, and obtained 270 samples in each fault condition, while 540 samples in normal condition (half of normal samples are used for training with multiple sources together). Each sample has 4000 points. Based on the eight domains, we conduct eight diagnosis tasks in our experiments, as listed in Table II.

B. Experimental Setup

1) Comparison Methods: In our experiments, several methods are compared as follows.
1) SVM: Support vector machine.
2) MLP: Multilayer perceptron.

3) KNN: k-nearest neighbor.
4) LR: Logistic regression.
5) TrAdaBoost [46].
6) SSTCA: Semisupervised transfer component analysis [47].
7) SGF: Sampling geodesic flow [35].
8) GFK: Geodesic flow kernel [34].
9) MSDGIFI: Our proposed multisource domain generalization method for Intelligent Fault Identification.

The first four methods are conventional machine-learning methods without transfer. TrAdaBoost and SSTCA are the representative single-source transfer learning methods and have been adopted to diagnosis problem [20], [21]. SGF and GFK are two subspace-based multisource transfer-learning methods that are close to our method.

2) Feature Extraction: We first extract 23 statistical parameters from the time and frequency domain to describe the health condition of the bearing [48]. In addition, the rotation frequency amplitude and four fault frequency amplitude in spectrum and envelope spectrum are also extracted [49]. There are 33 features in total. The formulas of each feature are listed in Table III. Let $\{a_i\}_{i=1,...,N}$ be a vibration signal, $\{\mathcal{F}(f_i)\}_{i=1,...,K}$ be the amplitude spectrum, and $\{\mathcal{E}(f_i)\}_{i=1,...,K}$ be the envelope spectrum of the signal. f_s is the rotation frequency, and $f_{\mathrm{BPFO}}, f_{\mathrm{BPFI}}, f_{\mathrm{BFF}}$, and f_{FTF} are the fault frequency of outer race, inner race, roller, and cage, respectively.

3) Implementation Details: Following [19], [50], we optimize parameter setting by searching the hyper-parameter space, and report the best results of each method. For conventional machine-learning methods, we use all samples from each source domain and normal samples of target domain together to train identification model. LIBSVM package is used to implement SVM [51]. Gaussian (RBF) kernel is adopted, and tradeoff parameter c is set to 1 throughout all tasks. This SVM classifier is the baseline and will be applied in SSTCA, SGF, GFK, and MSDGIFI as well. MLP has only one hidden layer, and the optimal number of neurons in the hidden layer hn is selected from $\{5, 10, 20, 40, 80\}$. Total 20 trials are carried out for each setting, and the best mean identification accuracy is chosen for further comparison. The maximum training epoch is 500, and the learning rate is 0.15. For KNN, we search the optimal number of nearest neighbor k in $\{1, 5, 9, 13, 17, 21, 25, 29, 33, 63\}$. For LR, the tradeoff parameter ρ is selected by searching $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. For TrAdaBoost, the base classifier is linear SVM, and the maximum number of iterations is 200. Total 20 trials are carried out for each diagnosis task, and mean accuracy is adopted for comparison. For SSTCA, we set regularization tradeoff parameter $\mu = 1$, supervise term tradeoff parameter $\gamma = 0.5$, and geometry term tradeoff parameter $\lambda = 1$. The optimal subspace dimension is selected by searching 2–30. We implement the experiments using TrAdaBoost and SSTCA in two different manners. One manner is to combine all source domains as a single source, and report the best diagnosis result using these two transfer-learning methods. The other one is to compute the average value of identification results achieved using each single source. For SGF, PCA subspace is used for each domain and three points are sampled from geodesic flow at $t = \{0, 0.5, 1\}$. The optimal subspace dimension is searched

TABLE III
FEATURES FOR ROLLING ELEMENT BEARING DIAGNOSIS

Features	Formula	Features	Formula	Features	Formula	Features	Formula
T_1	$\frac{1}{N}\sum_{i=1}^{N} a_i$	T_8	T_5/T_4	F_4	$\frac{1}{KF_2^2}\sum_{i=1}^{K}\left(\mathcal{F}_i - F_1\right)^4$	F_{11}	$\sum_{i=1}^{K}\left(f_i - F_5\right)^3 \mathcal{F}_i \big/ KF_6^3$
T_2	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(a_i - T_1\right)^2}$	T_9	T_5/T_3	F_5	$\sum_{i=1}^{K} f_i \mathcal{F}_i \big/ \sum_{i=1}^{K} \mathcal{F}_i$	F_{12}	$\sum_{i=1}^{K}\left(f_i - F_5\right)^4 \mathcal{F}_i \big/ KF_6^4$
T_3	$\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{|a_i|}\right)^2$	T_{10}	$T_4\big/\left(\frac{1}{N}\sum_{i=1}^{N}|a_i|\right)$	F_6	$\sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(f_i - F_5\right)^2 \mathcal{F}_i}$	F_{13}	$\mathcal{F}\left(f_s\right)$
T_4	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} a_i^2}$	T_{11}	$T_5\big/\left(\frac{1}{N}\sum_{i=1}^{N}|a_i|\right)$	F_7	$\sqrt{\sum_{i=1}^{K} f_i^2 \mathcal{F}_i \big/ \sum_{i=1}^{K} \mathcal{F}_i}$	$F_{14}-F_{17}$	$\left\{\begin{matrix}\mathcal{F}\left(f_{BPFO}\right), \mathcal{F}\left(f_{BPFI}\right)\\ \mathcal{F}\left(f_{BFF}\right), \mathcal{F}\left(f_{FTF}\right)\end{matrix}\right\}$
T_5	$\max|a_i|$	F_1	$\frac{1}{K}\sum_{i=1}^{K}\mathcal{F}_i$	F_8	$\sqrt{\sum_{i=1}^{K} f_i^4 \mathcal{F}_i \big/ \sum_{i=1}^{K} f_i^2 \mathcal{F}_i}$	E_1	$\mathcal{E}\left(f_s\right)$
T_6	$\sum_{i=1}^{N}\left(a_i - T_1\right)^3 \big/ (N-1)T_2^3$	F_2	$\frac{1}{K-1}\sum_{i=1}^{K}\left(\mathcal{F}_i - F_1\right)^2$	F_9	$\frac{\sum_{i=1}^{K} f_i^2 \mathcal{F}_i}{\sqrt{\sum_{i=1}^{K}\mathcal{F}_i \sum_{i=1}^{K} f_i^4 \mathcal{F}_i}}$	E_2-E_5	$\left\{\begin{matrix}\mathcal{E}\left(f_{BPFO}\right), \mathcal{E}\left(f_{BPFI}\right)\\ \mathcal{E}\left(f_{BFF}\right), \mathcal{E}\left(f_{FTF}\right)\end{matrix}\right\}$
T_7	$\sum_{i=1}^{N}\left(a_i - T_1\right)^4 \big/ (N-1)T_2^4$	F_3	$\frac{1}{K\left(\sqrt{F_2}\right)^3}\sum_{i=1}^{K}\left(\mathcal{F}_i - F_1\right)^3$	F_{10}	F_6/F_5	——	——

from $\{3, 5, 7, 9, 11, 13, 15\}$, note that $d < D/2$ for SGF. For GFK, we also use PCA subspace and report the best diagnosis result under single source. The optimal subspace is computed using subspace disagreement measure automatically [34]. SVM is used as classifiers for these two methods. For MSDGIFI, the optimal subspace dimension is searched from 2 to 30, and SVM classifier is also used. Referring to [37], we set $k = 7$ in all tasks. The optimal parameters for different methods can be found in the Appendix.

The identification accuracy on test data of the target domain is used to display the effectiveness, which is widely used in the literature [19], [47], [50]

$$Acc = \frac{|x : x \in \mathcal{D}_t \cap \tilde{y}_i = y_i|}{|x : x \in \mathcal{D}_t|}. \tag{9}$$

C. Performance Comparison and Analysis

1) Comparison of Diagnosis Results: The diagnosis results are listed in Table IV. From this table, it is seen that MSDGIFI achieves superior identification accuracies than four conventional machine-learning methods. Compared with the most competitive conventional method, i.e., KNN, the accuracy of MSDGIFI increases 4.61%. More importantly, the accuracy of our method increases 13.54% when compared with the baseline method, i.e., SVM. Moreover, our method improved the identification accuracies of SVM in each task.

Second, it is shown that the average identification result of our method is higher than single-source transfer learning methods. The accuracies of Multi_E, Multi_F, Multi_G, and Multi_H are promoted by our method remarkably. In fault diagnosis, the main issue of single-source transfer learning methods is how to choose an appropriate source for the target domain. In order to explain possible negative transfer in a single source scenario, we illustrate the transfer improvement (TI) [19] indicator of TrAdaBoost and SSTCA for single-source average experiments in Figs. 4 and 5. TI is computed by $TI = Acc\text{-}Acc_{\text{baseline}}$, where baseline is SVM trained using four sources. From the figures, it is shown that the diagnosis accuracy is not always increase when using different source domain to transfer. For example, in
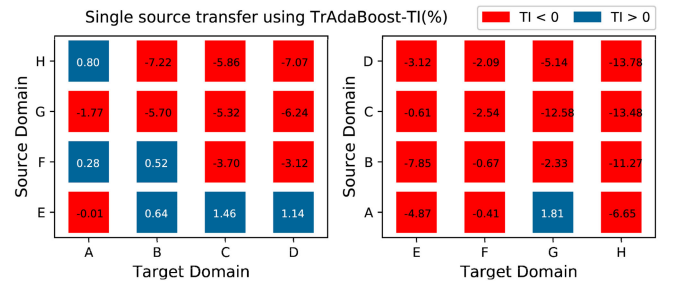

Fig. 4. TI of TrAdaBoost when diagnosing a target domain using a single source.
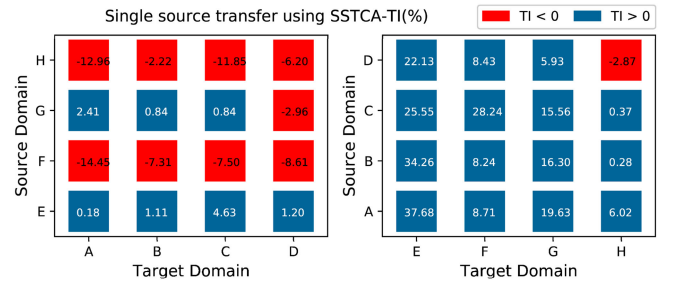

Fig. 5. TI of SSTCA when diagnosing a target domain using a single source.

Fig. 5 (SSTCA), when using H as source domain to diagnose B domain, the identification accuracy decrease 2.22% compared with baseline. Certainly, single-source transfer methods may achieve better results than MSDGIFI when using some given source domains, but our method shows superior domain generalization ability when considering multiple target tasks.

Besides, compared with SGF and GFK, our method also achieved superior identification performance, as shown in Table IV. The results above have confirmed the effectiveness of our method in bearing diagnosis problem. Through learning discriminant structure of each source using LFDA and finding mean subspace on Grassmann manifold, the proposed method could discover the general diagnosis knowledge and further promote generalization ability of classifier in target domain.

TABLE IV
IDENTIFICATION RESULTS (%) OF VARIOUS METHODS

Methods		Multi_A	Multi_B	Multi_C	Multi_D	Multi_E	Multi_F	Multi_G	Multi_H	Average
Conventional method	SVM	83.15	82.22	78.70	82.13	57.41	52.22	51.11	60.28	68.40
	MLP	82.74	81.89	76.87	82.41	58.33	61.41	52.15	55.46	68.91
	KNN	78.89	82.41	79.26	82.04	79.44	64.91	74.63	**77.04**	77.33
	LR	83.33	82.04	76.57	82.13	58.06	50.09	50.46	50.56	66.66
Single source Combine	TrAdaBoost	83.47	83.27	80.11	**83.33**	57.26	51.77	51.44	55.19	68.23
	SSTCA	**87.13**	83.33	**83.33**	**83.33**	82.59	62.87	57.50	58.70	74.85
Single source Average	TrAdaBoost	82.88	79.28	75.35	78.31	53.30	50.79	46.55	48.99	64.43
	SSTCA	76.95	80.33	75.23	77.99	87.32	65.63	65.47	61.23	73.77
Multisource transfer	SGF	79.69	81.50	76.41	81.81	64.17	59.98	50.09	57.71	68.92
	GFK	84.35	**83.52**	**83.33**	80.46	80.28	57.69	57.72	62.41	73.72
Our method	MSDGIFI	83.33	83.33	**83.33**	**83.33**	**89.81**	**81.57**	**77.41**	73.89	**81.94**

Therefore, our proposed method is potential for solving the actual fault diagnosis problem.

2) Empirical Analysis: To demonstrate the effectiveness of the proposed method intuitively, the distribution situation of samples in original space, SSTCA subspace, SGF space, and MSDG subspace has been visualized via *t*-SNE method [52]. *t*-SNE can visualize high-dimensional data by giving each data point a location in a 2-D or 3-D map. An indicator $Tc = (SI_s \cdot SI_t)/\exp(\mathrm{Dist}(\mathcal{D}_s, \mathcal{D}_t))$ is also constructed to measure transfer ability and class separability of several methods simultaneously and quantitatively. *Separability index* $SI = \sum_{i=1}^{n} I_{y_i}(y_{iN})/n$ is a class separability metric which is proposed by Thornton [53], where I is the indicator function with $I_{y_i}(y_{iN}) = 1$ if $y_i = y_{iN}$ and $I_{y_i}(y_{iN}) = 0$ if $y_i \neq y_{iN}$, y_i is the label of $\boldsymbol{x}_i$ and y_{iN} is the label of $\boldsymbol{x}_i$'s nearest neighbor. SI_s and SI_t are separability index of source domain and target domain, respectively. $\mathrm{Dist}(\mathcal{D}_s, \mathcal{D}_t) = \mathrm{MMD}(\mathcal{D}_s, \mathcal{D}_t)$ is the distribution distance between source domain and target domain, which is measured using maximum mean discrepancy (MMD) statistic [54] $\mathrm{MMD}(\mathcal{D}_s, \mathcal{D}_t) = \|1/n_s \sum_{i=1}^{n_s} \phi(\boldsymbol{x}_i^s) - 1/n_t \sum_{j=1}^{n_t} \phi(\boldsymbol{x}_j^t)\|_H$. The indicator $Tc \in (0, 1]$ and a larger Tc means that the samples of different classes are more separable or the distribution discrepancy between source and target domain is smaller.

Due to space limitation, we only illustrate the scatter plots and Tc values of Multi-F in Fig. 6. It is found that the samples of same category from different domains are much closer in MSDG space than in original space. Thus, the classifier trained using source domain data in MSDG space is more possible to identify the samples of target domain correctly. This demonstrates that the manner of discovering general diagnosis knowledge from multiple sources is favorable to improve the generalization ability of conventional data-driven diagnosis methods. Compared with SSTCA, although the distribution discrepancy between domains in MSDG space is larger, the samples of different categories in this space are more separable. It means that our manner of discovering common discriminant directions from multiple sources is more effective for fault identification than the manner of reducing distribution distance across domains, which is the main objective of SSTCA. Compared with SGF, it is found that the samples with same label from different domains are more
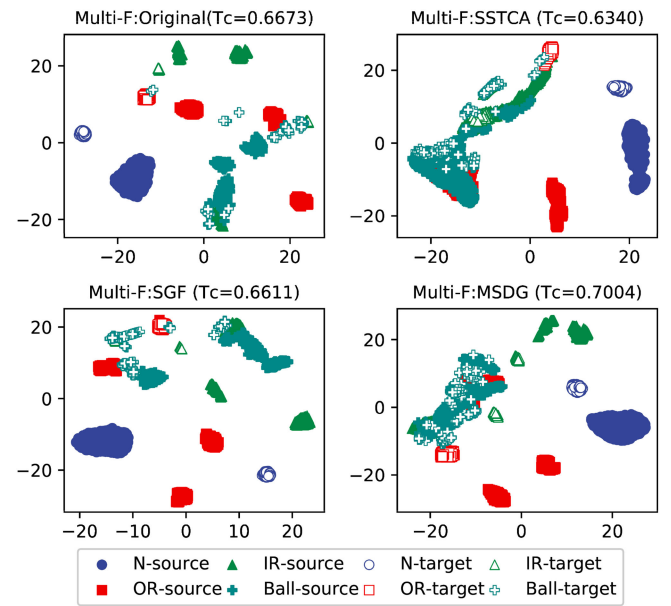

Fig. 6. Two dimensionality visualization via *t*-SNE under different spaces.

clustered in MSDG space. For example, the discrepancy of ball fault class (+ marker) from source and target domains is much smaller in MSDG space than in SGF space. The reason may be that SGF needs to embed target domain on Grassmann manifold. However, the embedded target domain on the manifold is not accurate due to unavailability of fault data in actual diagnosis scenario. Moreover, it is more separable between different classes in MSDG space. The reason may be that SGF uses PCA instead of LFDA to embed each domain on Grassmann manifold without considering the label information. These are also possible reasons that GFK cannot achieve accurate results in fault diagnosis applications.

D. Effectiveness of LFDA

In order to demonstrate the superiority of LFDA, we compared LFDA with other two dimensionality reduction methods: PCA and OLDA [38]. PCA is an unsupervised method
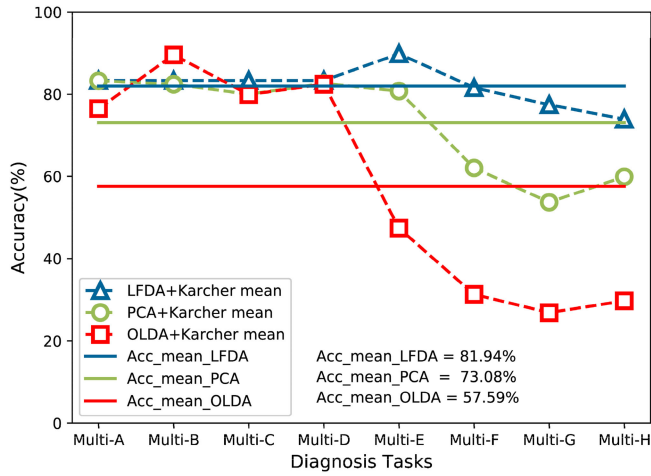
Fig. 7. Identification results using different subspace embedding methods.
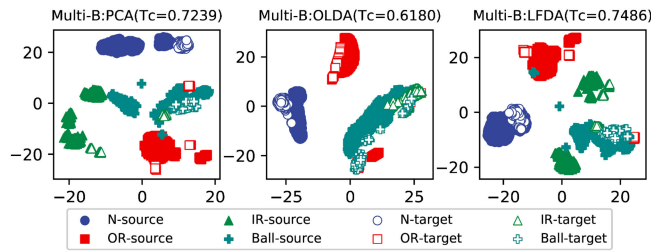

Fig. 8. Two dimensionality visualization via *t*-SNE under different subspace embedding methods (Multi-B task).
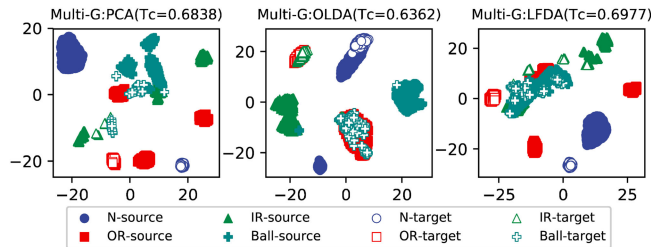

Fig. 9. Two dimensionality visualization via *t*-SNE under different subspace embedding methods (Multi-G task).

Fig. 10. Comparison of computational cost between different methods.

without considering label information, while OLDA is a supervised method without considering within-class local structure. We use different dimensionality reduction methods to implement subspace embedding and compute mean subspace in the same manner. The results of each diagnosis task are drawn in Fig. 7. We can observe that the mean identification accuracy when using LFDA is the highest. Moreover, the results of LFDA are more stable for different diagnosis tasks, which is a necessary characteristic for actual industry application.

As examples, 2-D visualization via *t*-SNE of Multi-B and Multi-G tasks under different subspace embedding methods is depicted in Figs. 8 and 9, respectively. From the figures, it is found that when learning subspace by LFDA, the samples of same category from different domains are much closer than using PCA. Thus, it is more favorable to implement health
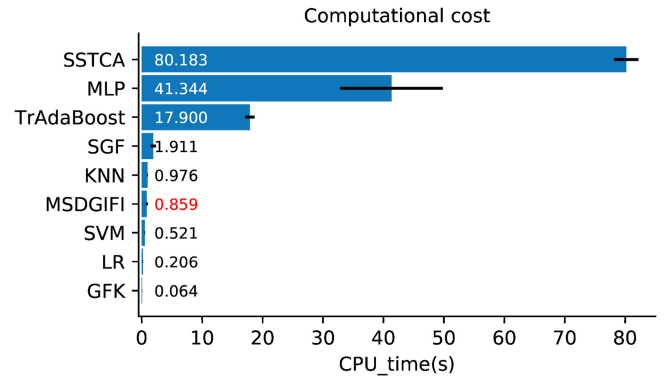
condition identification in LFDA space. Meanwhile, when using OLDA, although the samples of different classes are more separable, the distribution discrepancy between source and target domains is larger than using LFDA. This phenomenon is more obvious in the last four tasks. For example, the ball fault samples of target domain overlap with the outer race fault samples of source domains (the middle subfigure of Fig. 9). It means that the discriminant directions learned by OLDA from source domains cannot be generalized to a related target domain to implement effective fault identification. The reason may be that OLDA cannot learn effective discriminant directions when the distributions of source domains are multimodal or non-Gaussian. LFDA can reduce the dimensionality of multimodal data effectively through preserving local structure of data. In contrast to domain E–H, there are more clusters in domain A–D (see Fig. 3). Therefore, using LFDA as a subspace embedding method is more appropriate for Multi-E, Multi-F, Multi-G, and Multi-H, in which domains A–D are source domains. These results indicate that LFDA can discover more effective and stable mode discriminant directions than PCA and OLDA.

E. Comparison of Computational Cost

All methods in our experiments were conducted using MATLAB R2017b on a desktop computer equipped with an Intel Core i5-6402P processor (2.8 GHz), 8G RAM. For all methods, the average computational times of eight diagnosis tasks are displayed in Fig. 10. From this figure, it is found that the mean computational time of MSDGIFI is 0.859 s, and it is slightly longer than the SVM method. What is more, the computational efficiency of our proposed method is higher than several transfer learning methods, SGF, TrAdaBoost, and SSTCA.

F. Results Under Different Number of Sources

At last, we further discussed the influence of the number of sources to our method. For each diagnosis task, we add the sources from 2 to 7 step by step, and implement diagnosis using MSDGIFI method. It is worth to mention that for simulating the positive effect in considering more source domains, the more similar sources to target domain are added gradually. The identification results are depicted in Fig. 11. We also show the identification results of SVM model trained using the
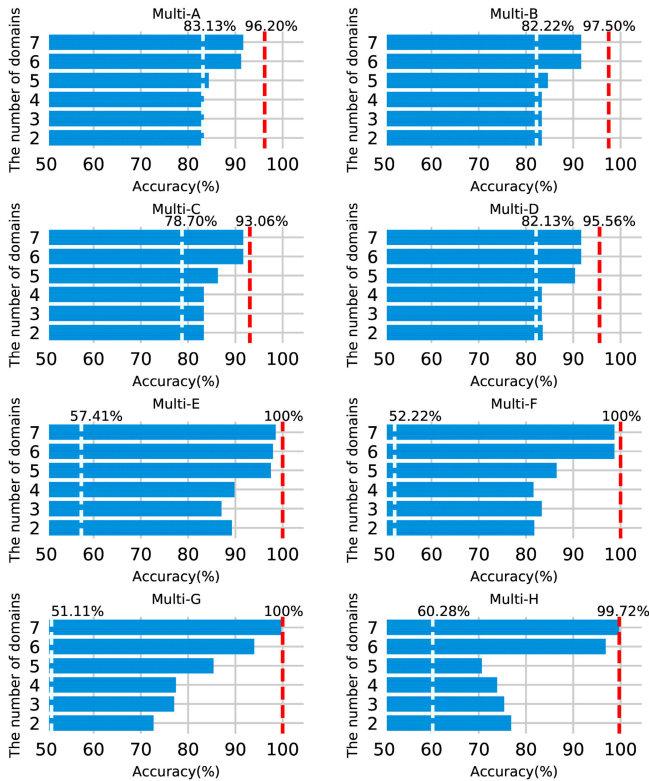
Fig. 11. Identification results under different number of source domains (red dashed lines show the identification results when using identically distributed data to train SVM model for target domain and white dashed lines show the identification results of SVM when the number of sources is four).

identically distributed data to target domain (red dashed lines in this figure). It can be seen that the identification accuracies increase with the number of sources and gradually approach the results of the model trained by identically distributed data. It indicates that our method has the potential to achieve significant result for actual diagnosis when considering more related sources. Besides, the identification results of baseline classifier SVM when using four sources are also displayed in Fig. 11 (white dashed lines). It is found that the proposed method even can achieve superior performance when just using two sources, which further verifies the effectiveness of MSDGIFI. There is an interesting phenomena in Multi-H that the identification accuracy decreases from 2 to 5 and increases under six sources. The reason may be that the distance between common subspace and the effective classification space of target domain becomes greater when source number changes from 2 to 5.

V. CONCLUSION

A new intelligent fault identification method based on multisource domain generalization (MSDGIFI) was proposed in this paper. MSDGIFI represented the fault discriminant structure of each source domain as a point on Grassmann manifold and finds the mean subspace through Karcher mean. The proposed method can discover effective and common diagnosis knowledge from multiple related source domains, and further

generalize the learned knowledge to new target tasks. Through considering multiple sources, the risk of negative transfer could be reduced in actual diagnosis scenario. In the diagnosis tasks of rolling element bearing, the proposed method achieved superior performance than other methods, and the effectiveness of our method has been verified and explained.

In the future work, we are going to study a multiple-domain feature evaluation technique to ensure the relevance between domains, and further explore the performance of our method in other fault diagnosis problem, such as gear fault diagnosis or diagnosing a run-to-failure bearing.

APPENDIX
OPTIMAL PARAMETERS FOR DIFFERENT METHODS

Throughout all tasks, $c = 1$ for SVM and maximum iteration number of TrAdaBoost is 200.

TABLE V
OPTIMAL PARAMETERS OF EACH METHOD IN EIGHT TASKS

Tasks	MLP	KNN	LR	SSTCA	SGF	GFK	MSDGIFI
Multi-A	$hn=80$	$k=9$	$\rho=10^{-3}$	$d=23$	$d=15$	F	$d=21$
Multi-B	$hn=80$	$k=1$	$\rho=10^{-3}$	$d=23$	$d=15$	F	$d=25$
Multi-C	$hn=10$	$k=9$	$\rho=10^{-3}$	$d=9$	$d=15$	F	$d=25$
Multi-D	$hn=80$	$k=9$	$\rho=10^{-3}$	$d=9$	$d=15$	F	$d=18$
Multi-E	$hn=10$	$k=9$	$\rho=10^{-1}$	$d=2$	$d=3$	A	$d=6$
Multi-F	$hn=80$	$k=17$	$\rho=10^{-3}$	$d=2$	$d=5$	A	$d=16$
Multi-G	$hn=5$	$k=13$	$\rho=10^{-3}$	$d=3$	$d=15$	B	$d=16$
Multi-H	$hn=10$	$k=1$	$\rho=10$	$d=2$	$d=15$	A	$d=17$
Common parameters	500; 0.15; sigmoid	--	--	$\mu=1$ $\gamma=0.5$ $\lambda=1$	$t=\{0, 0.5,1\}$	--	$k=7$

REFERENCES

[1] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657–667, Jan. 2015.

[2] Y. Wang, J. Xiang, R. Markert, and M. Liang, "Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications," *Mech. Syst. Signal Process.*, vol. 66/67, pp. 679–698, Jan. 2016.

[3] J. Chen, D. Zhou, C. Lyu, and C. Lu, "An integrated method based on CEEMD-SampEn and the correlation analysis algorithm for the fault diagnosis of a gearbox under different working conditions," *Mech. Syst. Signal Process.*, vol. 113, pp. 102–111, Dec. 2018.

[4] Y. Li, M. Xu, X. Liang, and W. Huang, "Application of bandwidth EMD and adaptive multiscale morphology analysis for incipient fault diagnosis of rolling bearings," *IEEE Trans. Ind. Electron.*, vol. 64, no. 8, pp. 6506–6517, Aug. 2017.

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

[6] D. Wang, "K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: Revisited," *Mech. Syst. Signal Process.*, vol. 70/71, pp. 201–208, Mar. 2016.

[7] M. Unal, M. Onat, M. Demetgul, and H. Kucuk, "Fault diagnosis of rolling bearings using a genetic algorithm optimized neural network," *Measurement*, vol. 58, pp. 187–196, Dec. 2014.

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

[9] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.

[10] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.

[11] M. S. Li, D. Yu, Z. Chen, K. Xiahou, T. Ji, and Q. H. Wu, "A data-driven residual-based method for fault diagnosis and isolation in wind turbines," *IEEE Trans. Sustain. Energy*, 2018, to be published, doi: 10.1109/TSTE.2018.2853990.

[12] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Trans. Ind. Inform.*, vol. 14, no. 7, pp. 3235–3243, Jul. 2018.

[13] J. Yu and Y. He, "Planetary gearbox fault diagnosis based on data-driven valued characteristic multigranulation model with incomplete diagnostic information," *J. Sound Vib.*, vol. 429, pp. 63–77, Sep. 2018.

[14] F. Cheng, J. Wang, L. Qu, and W. Qiao, "Rotor-current-based fault diagnosis for DFIG wind turbine drivetrain gearboxes using frequency analysis and a deep classifier," *IEEE Trans. Ind. Appl.*, vol. 54, no. 2, pp. 1062–1071, Apr. 2018.

[15] M. Zhao, M. Kang, B. Tang, and M. Pecht, "Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4290–4300, May 2018.

[16] Z. Liu, Z. Jia, C. Vong, S. Bu, J. Han, and X. Tang, "Capturing high-discriminative fault features for electronics-rich analog system via deep learning," *IEEE Trans. Ind. Inform.*, vol. 13, no. 3, pp. 1213–1226, Jun. 2017.

[17] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.

[18] R. Zhang, H. Tao, L. Wu, and Y. Guan, "Transfer learning with neural networks for bearing fault diagnosis in changing working conditions," *IEEE Access*, vol. 5, pp. 14347–14357, 2017.

[19] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep model based domain adaptation for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, Mar. 2017.

[20] J. Xie, L. Duan, and J. Wang, "On cross-domain feature fusion in gearbox fault diagnosis under various operating conditions based on transfer component analysis," in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, Ottawa, ON, Canada, 2016, pp. 1–6.

[21] F. Shen, C. Chen, R. Yan, and R. X. Gao, "Bearing fault diagnosis based on SVD feature extraction and transfer learning classification," in *Proc. Syst. Health Manage. Conf.*, Beijing, China, 2015, pp. 1–6.

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

[23] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual Domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[25] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, 2010, pp. 1855–1862.

[26] S. Sun, H. Shi, and Y. Wu, "A survey of multisource domain adaptation," *Inf. Fusion*, vol. 24, pp. 84–92, Jul. 2015.

[27] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Multisource domain adaptation and its application to early detection of fatigue," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 4, pp. 1–26, Dec. 2012.

[28] Z. Liu, J. Yang, H. Liu, J. Liu, and S. Member, "Learning from multiple sources via multiple domain relationship," *IEICE Trans. Inf. Syst.*, vol. 2, no. 7, pp. 1941–1944, 2016.

[29] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge transfer via multiple model local structure mapping," in *Proc. 14th ACM Int. Conf. Knowl. Discovery Data Mining*, Las Vegas, NV, USA, 2008, pp. 283–291.

[30] L. Duan, D. Xu, and I. W. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 1–15, Jan. 2012.

[31] T. Grubinger, A. Birlutiu, H. Schöner, T. Natschläger, and T. Heskes, "Multidomain transfer component analysis for domain generalization," *Neural Process. Lett.*, vol. 46, no. 3, pp. 845–855, Dec. 2017.

[32] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1414–1430, Jul. 2017.

[33] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," 2017, arXiv:1710.03463.

[34] B. Gong, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2066–2073.

[35] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 999–1006.

[36] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York, NY, USA: Academic, 1990.

[37] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[38] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *J. Mach. Learn. Res.*, vol. 6, pp. 483–502, Apr. 2005.

[39] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 376–383.

[40] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2008.

[41] F. Zhuang, X. Cheng, S. J. Pan, W. Yu, Q. He, and Z. Shi, "Transfer learning with multiple sources via consensus regularized autoencoders," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2014, vol. 8726, pp. 417–431.

[42] F. Zhuang, P. Luo, H. Xiong, Y. Xiong, Q. He, and Z. Shi, "Cross-domain learning from multiple sources: A consensus regularization perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 12, pp. 1664–1678, Dec. 2010.

[43] P. A. Absil, R. Mahony, and R. Sepulchre, "Riemannian geometry of Grassmann manifolds with a view on algorithmic computation," *Acta Appl. Math.*, vol. 80, no. 2, pp. 199–220, Jan. 2004.

[44] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a MAT-LAB toolbox for optimization on manifolds," *J. Mach. Learn. Res.*, vol. 15, pp. 1455–1459, Apr. 2014.

[45] K. Loparo, "Case Western Reserve University Bearing Data Center." [Online]. Available: http://csegroups.case.edu/bearingdatacenter/pages/download-data-file

[46] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007, pp. 193–200.

[47] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[48] Y. Lei, Z. He, and Y. Zi, "Application of an intelligent classification method to mechanical fault diagnosis," *Expert Syst. Appl.*, vol. 36, no. 6, pp. 9941–9948, Aug. 2009.

[49] R. B. Randall and J. Antoni, "Rolling element bearing diagnostics—A tutorial," *Mech. Syst. Signal Process.*, vol. 25, no. 2, pp. 485–520, Feb. 2011.

[50] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.

[51] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

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

[53] C. Thornton, "Separability is a learner's best friend," in *Proc. 4th Neural Comput. Psychol. Workshop*, London, U.K., 1997, pp. 40–47.

[54] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2006, pp. 513–520.

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

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

Rixin Wang received the B.E. and M.E. degrees in computer science and the Ph.D. degree in spacecraft design from the Harbin Institute of Technology, Harbin, China, in 1985, 1991, and 2003, respectively.

He is currently an Associate Professor with the Department of Engineering Mechanics, Harbin Institute of Technology. His research interests include fault detection and diagnosis for machinery and spacecraft.

Yuqing Li received the B.E. degree in mechanical design manufacturing and automation, the M.E. and Ph.D. degrees in general mechanics from the Harbin Institute of Technology, Harbin, China, in 2002, 2004, and 2008, respectively.

He is currently an Associate Professor with the Department of Engineering Mechanics, Harbin Institute of Technology, Harbin, China. His main research interests are planning and scheduling of spacecraft and spacecraft fault detection and diagnosis.
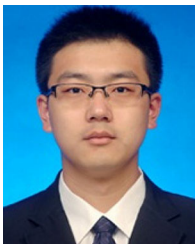
Yuantao Yang received the B.S. and the M.S. degrees in mechanics in 2013 and 2015, respectively, from the Harbin Institute of Technology, Harbin, China, where he is currently working toward the Ph.D. degree in mechanics with the Deep Space Exploration Research Center.

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

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

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