Coarse-to-Fine: Progressive Knowledge Transfer-Based Multitask Convolutional Neural Network for Intelligent Large-Scale Fault Diagnosis

Yu Wang, Ruonan Liu, *Member, IEEE*, Di Lin, *Member, IEEE*, Dongyue Chen, Ping Li, *Member, IEEE*, Qinghua Hu, *Senior Member, IEEE*, and C. L. Philip Chen, *Fellow, IEEE*

***Abstract*— In modern industry, large-scale fault diagnosis of complex systems is emerging and becoming increasingly important. Most deep learning-based methods perform well on small number of fault diagnosis, but cannot converge to satisfactory results when handling large-scale fault diagnosis because the huge number of fault types will lead to the problems of intra/inter-class distance unbalance and poor local minima in neural networks. To address the above problems, a progressive knowledge transfer-based multitask convolutional neural network (PKT-MCNN) is proposed. First, to construct the coarse-to-fine knowledge structure intelligently, a structure learning algorithm is proposed via clustering fault types in different coarse-grained nodes. Thus, the intra/inter-class distance unbalance problem can be mitigated by spreading similar tasks into different nodes. Then, an MCNN architecture is designed to learn the coarse and fine-grained task simultaneously and extract more general fault information, thereby pushing the algorithm away from poor local minima. Last but not least, a PKT algorithm is proposed, which can not only transfer the coarse-grained knowledge to the fine-grained task and further alleviate the intra/inter-class distance unbalance in feature space, but also regulate different learning stages by adjusting the attention weight to each task progressively. To verify the effectiveness of the proposed method, a dataset of a nuclear power system with 66 fault types was collected and analyzed. The results demonstrate that the proposed method can be a promising tool for large-scale fault diagnosis.**

***Index Terms*— Coarse-to-fine, knowledge transfer, large-scale fault diagnosis of complex system, multitask convolutional neural network (MCNN), structure learning.**

Manuscript received September 1, 2020; revised May 18, 2021; accepted July 17, 2021. This work was supported in part by the National Key Research and Development Project under Grant 2019YFB2101901 and Grant 2019YFB1703600; in part by the National Natural Science Foundation of China under Grant 61925602, Grant 61732011, and Grant 62006221; in part by the China Postdoctoral Science Foundation under Grant 2021TQ0242 and Grant 2021M690118; in part by the Innovation Foundation of Tianjin University under Grant 2021XZC-0066; in part by the Hong Kong Polytechnic University under Grant P0030419, Grant P0030929, and Grant P0035358; in part by the National Science and Technology Major Project under Grant 2017-I-0007-0008. *(Corresponding authors: Ruonan Liu; Qinghua Hu.)*

Yu Wang, Ruonan Liu, Di Lin, Dongyue Chen, and Qinghua Hu are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: ruonan.liu@tju.edu.cn; huqinghua@tju.edu.cn).

Ping Li is with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong.

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, also with the Navigation College, Dalian Maritime University, Dalian 116026, China, and also with the Faculty of Science and Technology, University of Macau, Macau 999078, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3100928.

Digital Object Identifier 10.1109/TNNLS.2021.3100928

I. Introduction

As an effective tool to keep the safe operation of industrial systems and reduce the unnecessary routine-shutdown maintenance costs, fault diagnosis has been increasingly significant in modern society [1], [2]. Therefore, a number of diagnosis methods have been proposed to detect faults early and accurately [3]–[5].

In recent years, with the development of sensor and information technology, the industrial data has been accumulated rapidly, which promotes the emergence of deep learning (DL)-based diagnosis methods [6]–[8]. Based on the deep architecture and multiple nonlinear layers, DL algorithms are able to learn high-level representation features adaptively and thus overcome inherent shortcomings of traditional diagnosis methods [9], [10].

The great contributions of these DL-based diagnosis methods are undeniable. But it can also be concluded from the literature that most state-of-the-art methods are component-specific and can perform well on classifying several fault types of a single or a few components. However, with the development of manufacturing industries, both the components and the faults are more and more complex and various, which leads to a number of fault types and large-scale fault diagnosis tasks. As a result, the vulnerabilities of DL-based methods are revealed when dealing with such large-scale diagnosis tasks with numerous fault types. First, DL-based methods starting with random initialization easily get stuck in a bad local minimum. Local minimum is the case that the gradient is close to zero but the points have a positive semidefinite Hessian, while a bad local minimum means the obtained local minimum is suboptimal to the a global minimum. It is pointed out by Erhan *et al.* [11] that increase of labels, that is, fault types, and complex structures will aggravate this problem. In addition, with the growing label space, the upper bound of the generalization error will increase and result in the decrease of final diagnosis performance [12]. Second, from the perspective of algorithm design, the increase of fault classes will lead to the problem of intra/inter-class distance unbalance [13]. For a certain class, the intra-class distance often refers to the Euclidean distance between samples within the class, while the inter-class distance refers to the Euclidean distance between samples of the class and samples of other classes [14]. Data with intra/inter-class distance unbalance have large intra-class distance and

small inter-class distance. For instance, the distances between similar failures of one component are small and hard to distinguish, while the distances between the failures in different subsystems are large and can be classified much easier. As a result, the inter-class distance of some similar failures can be even smaller than the intra-class distance of other failures in such large-scale fault diagnosis tasks, which makes traditional DL methods can hardly be applied. Therefore, compared with small-scale diagnosis problems, the large-scale fault diagnosis of complex systems with various fault types not only becomes vitally necessary, but also is a hard nut to crack.

The study in [15] has suggested complex systems often exhibit a hierarchical organization. When facing such a large-scale classification task, humans usually implement it along the hierarchical structure of candidate classes based on a coarse-to-fine strategy. For instance, when a vehicle is broken down, we tend to first determine the malfunction subsystem in a coarse-grained fault concept, such as the engine, the electrical system, or the fuel supply system. Then, the faulty component can be carefully diagnosed within the candidate breakdown subsystems. Such a hierarchical structure can be seen as a knowledge structure of fault types, which can provide additional and supplemental information in the coarse-grained for large-scale fault diagnosis.

In this article, such a knowledge structure is used to address the aforementioned problems in large-scale fault diagnosis for the following reasons: 1) the problem of poor local minima can be avoided by the macroscopical guidance of coarse-grained knowledge, because good initialization of both representations and classifiers is learned by convolutional neural networks (CNNs) in the coarse-grained task and 2) the problem of intra/inter-class distance change is alleviated by spreading similar faults into one coarse node and transferring the discriminant coarse-grained information to the fine-grained task, because the degree of this problem is greatly reduced in the coarse-grained task. Therefore, a progressive knowledge transfer-based multitask CNN (PKT-MCNN) framework is proposed based on a coarse-to-fine strategy for large-scale fault diagnosis of complex industrial systems. To obtain the knowledge structure automatically, a structure learning algorithm is also proposed to extract the coarse-to-fine knowledge structure by clustering fault types in different coarse-grained nodes of the structure. Then, a multitask CNN architecture is proposed to learn and transfer the extracted knowledge via a three-stage learning process, that is, coarse-grained task learning, multitask knowledge transfer, and fine-grained task fine-tuning, which provides a flexible and effective way to realize such a knowledge transfer process. A PKT algorithm is proposed and embedded in the multitask CNN to regulate different learning stages by paying different attention to each task dynamically, which can learn the coarse-grained knowledge and transfer it to the fine-grained task progressively.

The main contributions of this research are summarized as below.

1) This article proposes a novel coarse-to-fine diagnosis framework to make use of the knowledge structure in large-scale fault diagnosis. To adaptively and automatically extract the coarse-to-fine framework, a structure learning algorithm is also proposed. The experimental results verify that the learned nodes and hierarchical structure coincide with the physical composition of the diagnosed system.
2) Then, a multitask CNN architecture is designed to learn and transfer the extracted knowledge via a three-stage learning process, that is, coarse-grained task learning, multitask knowledge transfer, and fine-grained task fine-tuning, which provides a flexible and effective way to realize such a knowledge transfer process.
3) A knowledge transfer algorithm is proposed and embedded in the multitask CNN to regulate the different learning stages by paying different attention to each task dynamically and gradually transfer the attention from coarse- to fine-grained diagnosis task. Thus, the coarse-grained discriminant failure information can be learned and transferred to the fine-grained diagnosis task progressively.
4) A large-scale dataset of a nuclear power system was collected for experimental verification, which contains 362 995 samples from 66 fault types. The experimental results and comparisons with state-of-the-art diagnosis methods show that the proposed method can not only learn a logical coarse-to-fine structure, but also provide reliable diagnosis results for industrial big data, which highlights the effectiveness of the proposed method in large-scale fault diagnosis.

Therefore, the proposed framework satisfies the demand for large-scale fault diagnosis of complex systems and shows the potential to make the industrial systems more intelligent under such a big data environment. The code for this work is available at https://github.com/armstrongwang20/PKT-MCNN. Refer to the link for more details and reproduction.

The rest of this article is organized as follows. Section II illustrates the proposed framework in detail. In Section III, the proposed method is applied to analyze a large-scale dataset of a nuclear power system with 66 fault types. The results and comparisons with state-of-the-art methods verify the effectiveness of the proposed method in large-scale fault diagnosis. Finally, Section IV concludes this article.

II. Related Work

A. Fault Diagnosis Approaches

Traditional data-driven fault diagnosis methods first extract the fault features via signal processing algorithms such as Hilbert–Huang transform (HHT), wavelet packet transform, and sparse representation [16]. Then, the operational condition or the fault type can be recognized by machine learning methods, such as support vector machine (SVM) [17] and artificial neural network (ANN) [18]. However, due to the requirement of enough prior knowledge for feature extractor design and the limitation when facing the changeable fault types, traditional data-driven diagnosis methods become increasingly difficult to apply in modern industrial systems.

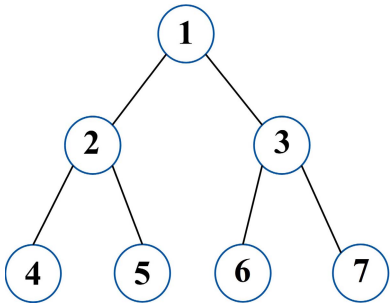


Fig. 1. Toy example of the coarse-to-fine structure.

Due to the rapid collection of industrial data, recent years witnessed the development of DL-based diagnosis methods, which can extract features automatically and adaptively. In literature, diverse DL models and their varieties have been applied for fault diagnosis successfully, including recurrent neural networks (RNNs) [19]–[21], sparse autoencoders (SAEs) [22], [23], deep belief networks (DBNs) [24], [25], and the highly popular CNNs [26]–[28]. Recently, to improve the generalization and the learning ability of target tasks, Chen *et al.* [29] proposed a transferable CNN to deal with the problem of insufficient training data, which has been investigated by different experimental datasets. Wang *et al.* [30] proposed a progressive optimized cascade CNN (C-CNN) structure to extract feature maps from different scales and to enable the algorithm to converge to a more optimum state. The effectiveness of the C-CNN has been verified by two motor fault diagnosis experiments. Azamfar *et al.* [31] proposed a 2-D CNN architecture for multisensor data fusion and gearbox fault diagnosis. The performance of the 2-D CNN has been evaluated by a motor current dataset obtained from a gearbox test rig.

However, when facing with a large number of fault types, traditional DL-based methods still have some limitations due to the intra/inter-class distance unbalance and local minima problem.

B. Coarse-to-Fine Learning

Coarse-to-fine structures contain fruitful relations information of different classes, which have been widely used for fine-grained classification and large-scale classification with many labels [13], [32], [33]. Deng *et al.* [32] used a semantic structure to divide a complex large-scale classification into several subproblems, and they designed a classification model that consists of multiple logistic regression (LR) classifier to label the samples from the coarse granularity, that is, root node, to the fine granularity, that is, the leaf node. Li *et al.* [34] leveraged the similar idea and applied it on few-shot learning. They integrated the structure into the learning phase of CNNs to learn multigranularity feature representations and then used the nearest-neighbor way to predict the test samples. Wei *et al.* [35] found that the coarse granularity of the structure can provide some complementary discriminative information to the fine granularity. Therefore, they refined the information of the learned coarse-grained features and attached them to the learning of the fine-grained task. Zhao *et al.* [13] transformed the coarse-to-fine structural information to several overlapped groups, which are applied to train multiple CNNs to collaborate for obtain accurate predictions in many-class recognition.

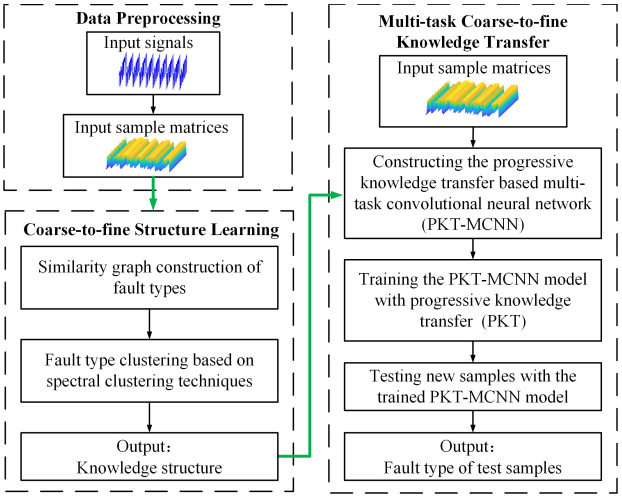


Fig. 2. Proposed framework.

Therefore, existing studies show that: 1) the coarse-grained labels in the structure have strong relations with fine-grained labels and 2) some representations and discriminative information for the coarse granularity can be helpful for the learning of the fine granularity. Therefore, we take advantage of the coarse-to-fine structure for CNNs to address the large-scale fault diagnosis problems. Fig. 1 shows a toy example of the coarse-to-fine structure, in which node #2 and node #3 are the coarse-grained nodes which are composed of fine-grained nodes #4, #5 and #6, #7, respectively.

III. Proposed Approach

A. Overview and Formulation

Knowledge structures usually exist in complex systems. For example, when there is a fault in a motor bearing, it can be regarded as a motor fault (higher level) or a bearing fault (lower level). Such knowledge can be beneficial for addressing the large-scale fault diagnosis task.

Therefore, a coarse-to-fine knowledge transfer framework is proposed for large-scale fault diagnosis, as shown in Fig. 2. First, the knowledge structure is extracted by composing various similar fault types into different coarse-grained fault concepts. Such knowledge is subsequently encoded in the parameters of the CNN by learning the coarse-grained task to obtain good initialization and coarse-grained information, which are then transferred to the learning of fine-grained task.

To extract the knowledge structure adaptively and automatically, a structure learning algorithm is proposed. Concretely, the similarity graph of fault types are first constructed (Section III-B1); then, the spectral clustering is applied to find the most similar groups of the fault types (Section III-B2).

Based on the extracted knowledge structure, a PKT-MCNN is proposed to learn and transfer the knowledge. First, a multitask CNN architecture is constructed to integrate the

processes of learning tasks in different granularity and transferring the coarse-grained knowledge to the fine-grained task (Section III-C). Then, a PKT algorithm is proposed to split the learning process into three stages, that is, coarse-grained task learning, multitask PKT, and fine-grained task fine-tuning, dynamically, by which different attention is given to different tasks adaptively based on the training process (Section III-D).

Formally, given training samples $\{\{\boldsymbol{x}_i^j\}_{i=1}^{M_i}\}_{j=1}^{N}$ in N fault types, where $\boldsymbol{x}_i^j$ is the jth sample in the ith fault type; M_i is the number of samples in the ith category; $i \in \{1, 2, \ldots, M_i\}$, and $j \in \{1, 2, \ldots, N\}$. The objective of the proposed framework is to extract the knowledge structure T, to train a PKT-MCNN which learns the coarse-grained task T_{C} and transfers the coarse-grained knowledge, for example, useful features and discriminant information, to the fine-grained fault diagnosis task T_{F}.

B. Coarse-to-Fine Structure Learning

1) Similarity Graph Construction of Fault Types: Knowledge structures usually exist in data, but they are not always explicitly available in many learning tasks. Moreover, it is difficult to construct such structures due to the incomplete physical structure information or a lack of domain human experts. To extract the coarse-grained knowledge structure automatically, the information in data has been used to group similar fault types into a common superordinate node of the structure, which represents a coarse-grained fault concept.

Given the jth sample in the ith fault type $\boldsymbol{x}_i^j$, the similarity graph $G = (V, E)$ contains the similarity information for each two fault types, where each vertex $\boldsymbol{v}_i \in V$ is a fault type and each edge $e_{ij} \in E$ is the degree of similarity between the two connected vertices $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$. To build this graph, the similarity of each pair of fault types should be calculated. Different from instance-level clustering, each fault type is required to be represented before computing the similarity. In this work, each fault type is vectorized and represented by the centroid of all its samples. Mathematically, a vertex $\boldsymbol{v}_i$ in the similarity graph is represented as

$$\boldsymbol{v}_i = \frac{1}{M_i} \sum_j \boldsymbol{x}_i^j \tag{1}$$

where M_i is the number of samples of the ith fault type. There are two advantages of this representation method for large-scale fault diagnosis. On the one hand, it is efficient to compute the first-order statistic, that is, the mean vector, for many fault samples. On the other hand, representing a fault type by the mean vector can alleviate the adverse influence of noisy samples, which usually exist in large-scale dataset. With the representation of vertices, the similarity information e_{ij} between each two fault types i and j can be computed by the Gaussian similarity

$$e_{ij} = \exp\left(-\frac{||\boldsymbol{v}_i - \boldsymbol{v}_j||^2}{(2\sigma)^2}\right) \tag{2}$$

where σ is the scaling factor that normalizes the value of e_{ij} to [0, 1].

2) Fault Type Clustering: With the similarity graph G, the fault types can be assigned to different coarse-grained fault concepts through clustering techniques. In this article, the normalized cut (NCut) algorithm is applied to cut the graph into several subgraphs, which can be formalized to optimize the following objective function:

$$\min_{C_1, C_2, \ldots, C_k} \mathrm{Tr}\left(H'LH\right) \quad \text{s.t. } H'DH = I \tag{3}$$

where k is the number of coarse-grained fault concepts to be clustered;

$$h_{ij} = \begin{cases} 1/|C_j|, & \text{if } v_i \in C_j \\ 0, & \text{otherwise} \end{cases}$$

is the element in the matrix H, $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, k\}$, $L = D - G$ is the Laplacian matrix, D is the degree matrix of G, I is the identity matrix, and $|\cdot|$ is the cardinality of a set. Shi and Malik [36] showed that this objective can be approximated by the eigenvector of L associated with the second smallest eigenvalue. In this way, k coarse-grained fault concepts $\{C_1, C_2, \ldots, C_k\}$ are obtained to form a two-level knowledge structure T, in which each coarse-grained concept contains several fine-grained fault types.

C. Multitask Convolutional Neural Network

To learn and transfer the coarse-grained knowledge to the fine-grained task, a multitask CNN architecture is designed by sharing the representation learning layers and owning corresponding task-specific learning layers. A multitask CNN is an inductive transfer method that uses the domain-specific information contained in the training signals of related tasks by learning the multiple tasks in parallel while using a shared representation and thus improves learning for one task by using the information contained in related tasks [37]–[39]. The merits of this method are twofold: on the one hand, the learning processes of the coarse- and fine-grained tasks are unified in an integrative method, which makes the method take advantage of related tasks in different granularity in an end-to-end optimization way; on the other hand, the shared representation learning layers act as a medium to store and fuse the knowledge in different grained tasks, which enables the transfer of the knowledge and helps the large-scale diagnosis tasks to be learned more effectively.

1) Shared Representation Learning Layers: Typically, representation learning layers often consist of convolutional layers and pooling layers in a CNN. For a given sample $\boldsymbol{x}$, each kernel is convolved across the width and height of $\boldsymbol{x}$, which computes the dot product between the kernel and $\boldsymbol{x}$. The max-pooling is applied in this article, which chooses the maximum value within a pooling region and propagates to the next layer. The representation learning layers often stack multiple convolutional layers and pooling layers, denoted as

$$\boldsymbol{W}_r = \left[\left[\boldsymbol{W}_1; \ldots; \boldsymbol{W}_{K_1}\right]; \ldots; \left[\boldsymbol{W}_1; \ldots; \boldsymbol{W}_{K_m}\right]\right] \tag{4}$$

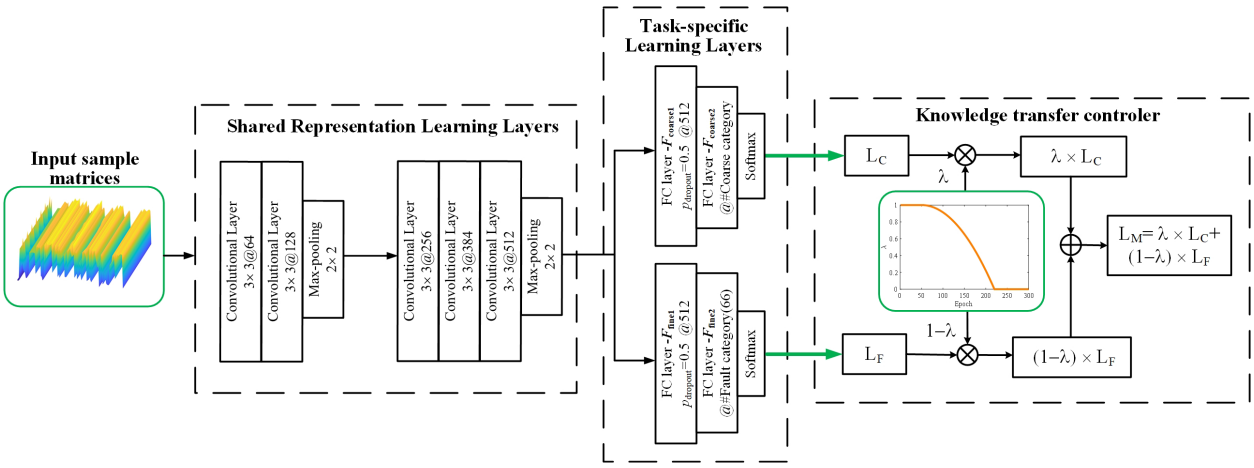


Fig. 3. PKT-MCNN framework, where L_{C}, L_{F}, and L_{M} represent the loss of coarse-grained, fine-grained, and multigrained task, respectively.

where m is the number of convolutional layers, and $K_i (i \in \{1, \ldots, m\})$ is the number of convolutional kernels in the ith convolutional layer.

2) Task-Specific Learning Layers: Based on the learned representations, a fully connected (FC) layer consists of a simple multilayer perceptron (MLP) that learns to weight the representations to identify the object fault type. Subsequently, a softmax function is performed to map the logits produced by the FC layer to [0, 1]. Mathematically, it can be written as

$$\boldsymbol{y} = \mathrm{softmax}\big(\boldsymbol{W}_f(\boldsymbol{Z}_r) + b\big) \tag{5}$$

where $\boldsymbol{y}$ is the output of the FC layer with softmax function, $\boldsymbol{W}_f$ is the weight vector of the FC layer, b is the bias term, and $\boldsymbol{Z}_r$ is the output of the shared representation learning layers. In this work, there are two tasks to be learned: the coarse-grained task T_{C} and the fine-grained task T_{F}. In this regard, each task is attached with a specific FC layer to learn different discriminative information, denoted as $\boldsymbol{W}_f^{\mathrm{C}}$ and $\boldsymbol{W}_f^{\mathrm{F}}$, for the coarse-grained task and the fine-grained task, respectively.

D. PKT Algorithm

To make the multitask CNN pay attention to the learning on different tasks for effective transfer, a novel PKT algorithm is designed and embedded on the top of the multitask CNN. The framework of PKT-MCNN is shown in Fig. 3 (the displayed CNN structure is model 12 in Table I). Concretely, the PKT operates on the loss layers of the multitask CNN and splits the training process of the multitask CNN into three stages: coarse-grained task learning, multitask PKT, and fine-grained task fine-tuning. The PKT controls and switches the stages through the weights of the loss layers, which represent the current degree of attention to the corresponding tasks. The loss function used here is the cross-entropy loss, denoted as

$$\mathcal{L}\big(\boldsymbol{W}_f|\boldsymbol{x}\big) = -\sum_{i=1}^{N} r_i \log p_i\big(\boldsymbol{W}_f|\boldsymbol{x}\big) \tag{6}$$

where r_i is 0 or 1 corresponding to the true label, $p_i(\cdot)$ producing the scores of the prediction, and it is the function of the learnable parameter $\boldsymbol{W}_f$. Consequently, the loss function $\mathcal{L}_{\mathrm{M}}$ of PKT-MCNN is defined as the weighted combination of the coarse-grained and fine-grained tasks

$$\begin{aligned} \mathcal{L}_{\mathrm{M}} &= \lambda\mathcal{L}_{\mathrm{C}} + (1-\lambda)\mathcal{L}_{\mathrm{F}} \\ \mathcal{L}_{\mathrm{C}} &= \mathcal{L}\big(\big[\boldsymbol{W}_r; \boldsymbol{W}_f^{\mathrm{C}}\big]|\boldsymbol{x}\big) \\ \mathcal{L}_{\mathrm{F}} &= \mathcal{L}\big(\big[\boldsymbol{W}_r; \boldsymbol{W}_f^{\mathrm{F}}\big]|\boldsymbol{x}\big) \end{aligned} \tag{7}$$

where λ is the weight that accounts for the learning attention of the two tasks. Recall that there are three main stages in the training process of the proposed PKT-MCNN: 1) the coarse-grained task is trained to grasp the coarse-grained knowledge; 2) the coarse- and fine-grained tasks are trained simultaneously in a multitask way, where the method can not only make the two tasks benefit each other, but also progressively transfer the obtained coarse-grained knowledge to the fine-grained task; and 3) the fine-grained task is trained individually by fine-tuning the parameters that have been updated in stage 2) to have a more accurate understanding of the current task.

In this respect, the PKT algorithm is proposed to control and switch different learning stages by adjusting the attention to different tasks. It is shown by Cipolla *et al.* [40] that the weight of the loss term has a great impact on the learning process, where the task with a large loss value would update more quickly than that with a small loss value. Therefore, the PKT regulates the different attention to the two tasks via a weight parameter λ. Specifically, in the first stage, λ is set to be 1 to learn the coarse-grained task only, because the weight of the fine-grained task is 0 and the loss is stopped from back-propagating. In the second stage, λ gradually decreases until 0 according to the number of the training epoch, and this aims to learn both tasks simultaneously and progressively transfer the coarse-grained knowledge to the fine-grained one. Concretely, λ in the second stage, denoted as λ_2, is computed by

$$\lambda_2 = 1 - \left(\frac{B - B_1}{B_{\max} - B_1 - B_3}\right)^2 \tag{8}$$

where B_1 and B_3 are the number of training epochs in stages 1) and 2), respectively, B is the number of the current epoch, and $B_{\max}$ is the maximum number of epochs. The reason of

designing the function (8) for λ here is that the attention to the fine-grained task is supposed to increase rapidly once the coarse-grained task is well trained. Finally, the value of λ in the third stage is contrary to that in the first stage. λ is set to 0 to fine-tune the fine-grained task without updating the parameters of the coarse-grained task.

The training procedure is summarized in Algorithm 1. First, the coarse-grained fault concepts are extracted to form the knowledge structure, by which the similarity graph is constructed before clustering the fault types into several groups. Subsequently, a multitask CNN is designed to learn and transfer the coarse-grained knowledge. Then, the proposed PKT algorithm regulates the training stage by allocating different loss weights to different tasks via parameter λ, which is set to be 0, 1, and $1-((B-B_1)/(B_{\max}-B_1-B_3))^2$, respectively. The testing procedure is the same with conventional CNNs, where the algorithm predicts the input sample as one of the fault types without considering the coarse-grained branch.

The proposed method is general in cases that consist of many classes. For image data, despite the existence of large-scale pre-training models, the proposed method can still benefit the training of the model in terms of finding a better local minimum. For the structure of the proposed method, FC layers and a classifier are attached to each task in the fault diagnosis problem. For image data, it may need simple adjustment by adding additional convolutional layers for each task before the FC layer according to the characteristics of data. Additionally, the proposed method can also be applied to structures with more than two levels by transferring the high-level knowledge to the low level layer by layer.

Algorithm 1 Training Procedure of the Proposed Approach

Input: Training samples $\{\{\boldsymbol{x}_i^j\}_{i=1}^{M_i}\}_{j=1}^{N}$.
Output: Knowledge structure $\boldsymbol{T}$, multitask CNN $\boldsymbol{W}=[\boldsymbol{W}_r;\boldsymbol{W}_f^C;\boldsymbol{W}_f^F]$.
Extract the knowledge structure;
1. Construct the similarity graph G according to Equation (1) and (2);
2. Build the knowledge structure T through clustering coarse-grained fault concepts by optimizing the objective (3);
Transfer knowledge through training the multitask CNN;
3. Update the model $\boldsymbol{W}=[\boldsymbol{W}_r;\boldsymbol{W}_f^C;\boldsymbol{W}_f^F]$ through optimizing loss (7) in first B_1 epochs with $\lambda=1$ produced by the PKT algorithm;
4. Update the model $\boldsymbol{W}=[\boldsymbol{W}_r;\boldsymbol{W}_f^C;\boldsymbol{W}_f^F]$ through optimizing loss (7) in $(B_{max}-B_3-B_1)$ epochs with λ generated based on Equation (8) by the PKT algorithm;
5. Update the model $\boldsymbol{W}=[\boldsymbol{W}_r;\boldsymbol{W}_f^C;\boldsymbol{W}_f^F]$ through optimizing loss (7) in last B_3 epochs with $\lambda=0$ produced by the PKT algorithm;
return $\boldsymbol{T}$, $\boldsymbol{W}$.

IV. Experimental Study

Currently, most existing datasets usually consist of a few fault type, but users may encounter dozens of possible fault classes in real industrial applications. In this regard, a new large-scale dataset is collected and analyzed by the proposed method, which is described in Section IV-A, namely FAult Recognition Of Nuclear power system (FARON), which contains 362 995 samples from 66 types. In Sections IV-B and IV-C, the effectiveness of the proposed approach is verified by conducting the large-scale diagnosis task based on this new dataset in comparison with 12 baseline CNN models. The ablation study is carried out in Section IV-F, and the analyses on parameters and the extracted knowledge are discussed in Sections IV-G and IV-H, respectively.

A. Data Description

The FARON dataset consists of many encrypted operational data of a nuclear power system, which is composed of a reactor, a main coolant pump, a pressurizer, a steam generator, a feed pump, feed water heaters, a condensate pump, a sea water pump, a separator, and a reheater, as shown in Fig. 4. Different conditions were simulated by a nuclear power system simulator with 121 sensor-respond outputs for the primary system and the secondary system. Data under normal operational environment were collected by 5.32-h simulation. During the process of fault data generation, the nuclear power plant started from the normal state in each simulation and ran for 2 min; then the faults were introduced at a certain point in the operational process. Thus, the operational data of 65 fault types were collected by 19.89-h simulation, in which the simulation time of each fault type ranges from 10 to 77.7 min. There are 76 632 samples under health state and 286 363 samples of different faults, respectively. Six raw signal examples of the feed water pump and the main condenser under different health conditions are displayed in Fig. 5. In order to extract the spatial and temporal domain features, the collected time-series multivariable samples were cut into time segments to form a 2-D $m \times n$ matrix, where m is the length of sampling time and n is the number of variables. In this article, m is 20 according to [41] and n is 121. 70%, 10%, and 20% sample matrices are selected according to the chronological order as the training set, the validation set, and the test set, respectively. To simulate the actual conditions, followed by [42], the raw

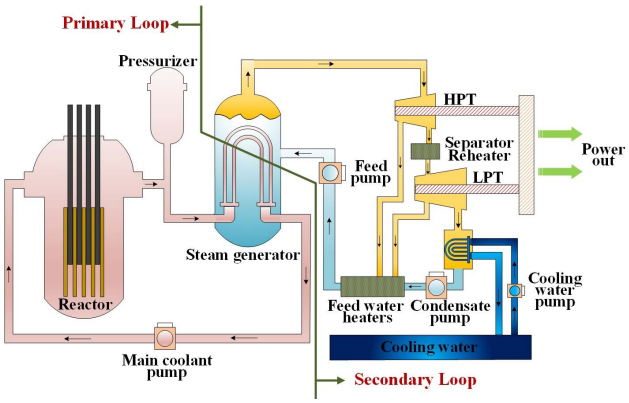


Fig. 4. Experimental setup of the nuclear power system.

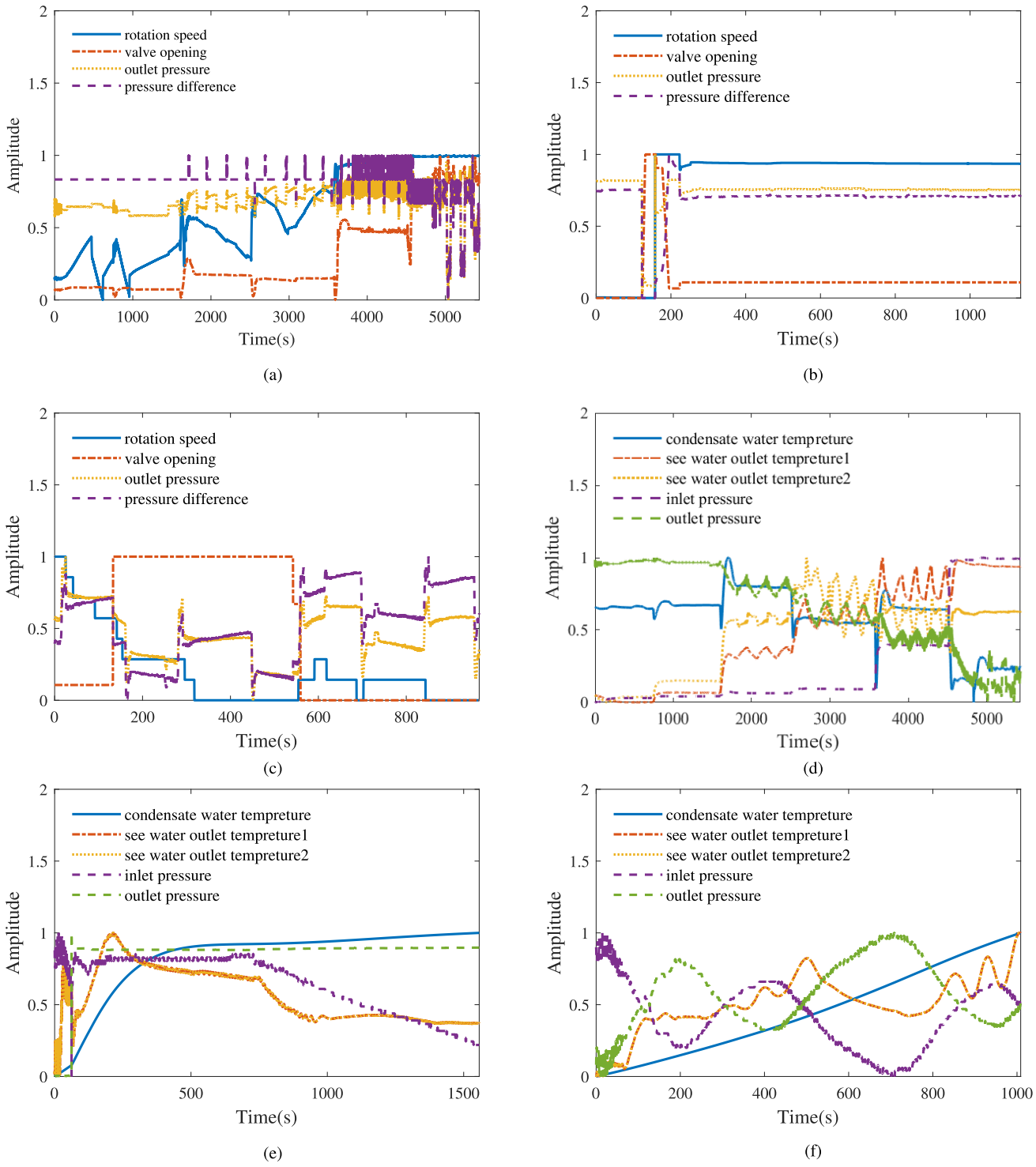


Fig. 5. Six raw signal samples under different health conditions: (a) health feed water pump; (b) feed water pump fault; (c) valve fault; (d) health main condenser; (e) main condenser pump fault; and (f) valve fault.

data are added by white noise whose mean value is 0 and the standard deviations are randomly assigned from 0 to 0.05 for different dimensions of the raw data.

B. Experimental Setups and Implementations

1) Experimental Setups: To verify the effectiveness of the proposed approach comprehensively, 12 CNNs with different architectures are adopted in this study as the baselines. The reasons are mainly twofold: 1) the proposed method is designed based on the conventional CNN model and aims to improve it and 2) different architectures of CNNs are supposed to be compared empirically to eliminate the bias of the CNN architecture. Concretely, the architectures of different CNN models used in this study are shown in Table I, and they are adjusted based on CNN structures used in previous studies [41], [43]. In this article, classification accuracy (A)

TABLE I
STRUCTURE OF CNNS

# Model	Structure
Model 1	Conv(128)-Maxpool-FC(512)*-FC(66)
Model 2	Conv(128)-Conv(256)-Maxpool-FC(512)*-FC(66)
Model 3	Conv(128)-Conv(256)-Conv(384)-Maxpool-FC(512)*-FC(66)
Model 4	Conv(128)-Conv(256)-Conv(384)-Conv(512)-Maxpool-FC(512)*-FC(66)
Model 5	Conv(64)-Maxpool-Conv(192)-Maxpool-FC(512)*-FC(66)
Model 6	Conv(64)-Conv(64)-Maxpool-Conv(256)-Maxpool-FC(512)*-FC(66)
Model 7	Conv(64)-Conv(128)-Maxpool-Conv(256)-Maxpool-FC(512)*- FC(66)
Model 8	Conv(64)-Conv(128)-Maxpool-Conv(256)-Maxpool(1×1)-FC(512)*-FC(66)
Model 9	Conv(64)-Conv(192)-Conv(320)-Maxpool-FC(512)*-FC(66)
Model 10	Conv(64)-Conv(128)-Maxpool-Conv(256)-Conv(384)-Maxpool-FC(512)*-FC(66)
Model 11	Conv(64)-Conv(128)-Conv(192)-Maxpool-Conv(320)-Conv(448)-Maxpool-FC(512)*-FC(66)
Model 12	Conv(64)-Conv(128)-Maxpool-Conv(256)-Conv(384)-Conv(512)-Maxpool-FC(512)*-FC(66)

FC*: Dropout ($p = 0.5$) is used for this FC layer.
The default kernel size of Conv layer is set to 3×3.
The default kernel size of Maxpooling layer is set to 2×2.

is used as evaluation indexes to compare the performance of different methods. To be specific,

$$A = \frac{\sum_i (\mathrm{TP}_i + \mathrm{TN}_i)}{\sum_i (\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{TN}_i + \mathrm{FN}_i)} \tag{9}$$

where TP is the true positive, TN is the true negative, FP is the false positive, FN is the false negative, and i is the fault type.

2) Implementation Details: In the training process of all the methods, including 12 baseline CNNs and their improved version PKT-MCNNs, the Adam optimization algorithm was applied, in which the learning rate was set as 10^{-5}, and other parameters were set by default. The batch size was set to 256. The number of epochs for coarse-grained knowledge task learning, multitask PKT, and fine-grained task fine-tuning were set to 100, 50, and 150, respectively. σ in (2) was set to 0.01 All the diagnosis methods were implemented by Pytorch, and the experiments were carried out by a high-performance server with 60 GB memory and 4 GeForce RTX-2080 GPUs. The average of ten trials of each diagnosis method is reported in this article as the final result.

C. Results on Fault Diagnosis Task

The performance of the proposed method is compared with common used fault diagnosis methods. Six methods are performed on the dataset for comprehensive comparison, including k-nearest neighbors (KNN), SVM, LR, MLP, DBN, and CNN, which are used in many previous studies [41], [43]. In this experiment, the structure of CNN and the proposed method uses model 1. The results are shown in Table II. It can be seen that the proposed method outperforms other methods in a large margin, which verifies its effectiveness. Moreover, the accuracies of both baseline CNN and the corresponding PKT-MCNN models with all 12 structures on raw data and

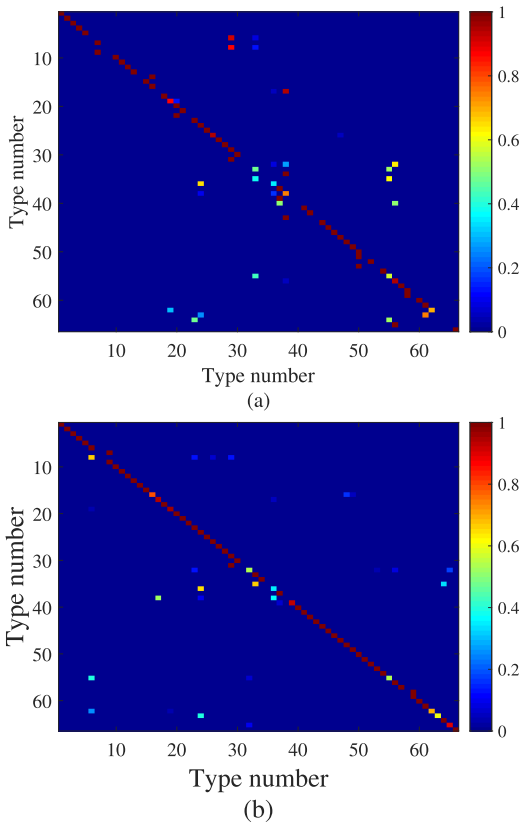


Fig. 6. Confusion matrix of (a) conventional flat CNN and (b) PKT-MCNN with model 1. The rows correspond to the predicted class (output class) and the columns correspond to the true class (target class).

TABLE II
ACCURACY (%) COMPARISON BETWEEN DIFFERENT FAULT DIAGNOSIS METHODS WITH RAW INPUT AND DATA WITH WHITE NOISE

Input	KNN	Flat CNN	SVM	LR	MLP	DBN	PKT-MCNN
Raw	74.63	82.47	79.12	76.53	63.12	64.94	**88.42**
WN	70.19	78.92	76.87	72.06	60.55	61.19	**85.30**

data with noise are shown in Tables III and IV. It can be seen that the proposed framework shows clear advantages for large-scale fault diagnosis over the baseline CNNs on all 12 different tested architectures both on the raw data and actual noisy data. This is because the good initialization of CNN parameters obtained from the coarse-grained task can effectively avoid the poor local minima. Moreover, useful discriminant information is retained and transferred to the fine-grained task for effective fault identification. The confusion matrices of the flat CNN and the proposed PKT-MCNN are also displayed in Fig. 6 (with model 1) and demonstrate the effectiveness of the proposed method. In addition, the coarse-grained task can be regarded as the direct upstream task for the objective fine-grained fault diagnosis and thus helps the CNN learn some useful features and discriminant information that are difficult to learn from the fine-grained task directly. Fig. 7 shows the training and testing processes of PKT-MCNN and flat CNN with model 5. It can be seen that the accuracy curve of PKT-MCNN keeps increasing after the intersection point of curves of the flat CNN and the PKT-MCNN, which means that PKT-MCNN converges to a better local minima than the flat

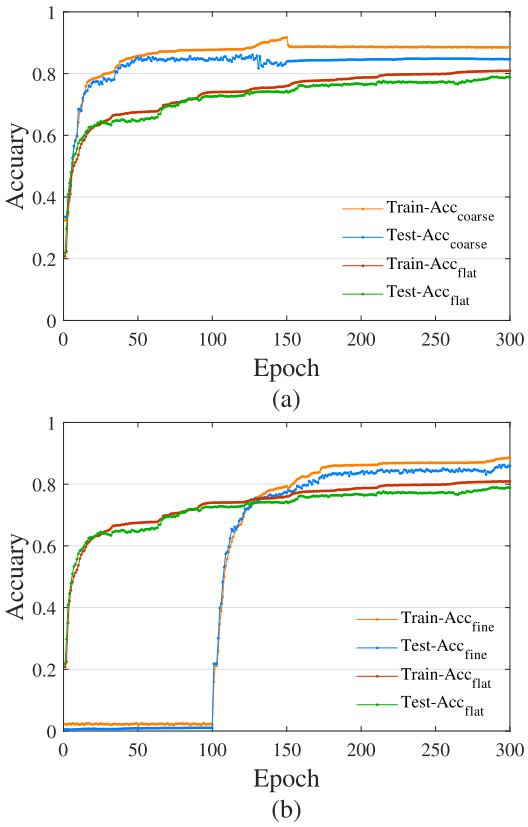


Fig. 7. Accuracies for the iteration process of the PKT-MCNN and the flat CNN: (a) coarse-grained task accuracy and (b) fine-grained task accuracy, and accuracy of flat CNN is plotted in both subfigures.

TABLE III

ACCURACY (%) COMPARISON BETWEEN DIFFERENT KNOWLEDGE TRANSFER-BASED METHODS ON RAW DATA

# Model	Flat CNN	KD	AB-KD	PKT-MCNN
Model 1	79.92	80.91	81.22	**85.30**
Model 2	78.77	81.62	82.17	**85.14**
Model 3	79.28	80.01	80.91	**85.11**
Model 4	75.40	77.71	78.35	**80.14**
Model 5	78.13	80.81	81.48	**82.78**
Model 6	76.51	79.08	81.00	**82.92**
Model 7	76.99	79.25	80.90	**82.71**
Model 8	76.31	78.29	79.51	**82.33**
Model 9	76.29	79.96	80.48	**81.92**
Model 10	74.11	77.97	78.05	**79.91**
Model 11	73.25	77.03	78.21	**81.52**
Model 12	72.05	75.14	76.04	**78.11**

CNN. This shows that the PKT has a significant influence on the learning of CNNs [Fig. 7(b)]. Moreover, the figure shows that the multitask learning stage (epoch 100–150) would make the method retain the ability for the coarse-grained task and improve the ability for the fine-grained task. In the fine-grained fine-tuning stage (stage 3, after epoch 150), the performance of the coarse-grained task gradually decreases until convergence, while that of the fine-grained task keeps increasing. This result verifies that the three stages are all important for the knowledge transfer process.

To illustrate the learned essential features graphically, the t-distributed stochastic neighbor embedding (t-SNE) method [44] is employed to provide 2-D visual representations

TABLE IV

ACCURACY (%) COMPARISON BETWEEN DIFFERENT KNOWLEDGE TRANSFER-BASED METHODS ON DATA WITH WHITE NOISE

# Model	Flat CNN	KD	AB-KD	PKT-MCNN
Model 1	82.47	82.78	83.57	**88.42**
Model 2	82.32	84.37	83.00	**87.45**
Model 3	83.11	83.80	83.86	**87.09**
Model 4	79.88	82.09	81.80	**82.94**
Model 5	81.04	82.89	84.51	**85.23**
Model 6	80.23	83.51	84.14	**85.86**
Model 7	80.38	83.51	84.83	**85.57**
Model 8	81.25	84.60	82.60	**85.28**
Model 9	82.43	82.66	83.74	**84.71**
Model 10	78.92	82.77	81.97	**82.89**
Model 11	79.98	82.43	79.75	**83.32**
Model 12	77.73	80.57	79.24	**81.29**

TABLE V

ACCURACY (%) COMPARISON BETWEEN THE PROPOSED METHOD WITH DIFFERENT TRAINING SCHEMES AND CLUSTERING METHODS

# Model	Flat CNN	PKT-MCNN	CF-MCNN	CM-MCNN	KM-PKT
Model 1	82.47	**88.42**	85.33	83.11	88.27
Model 2	82.32	**87.45**	85.59	81.53	87.02
Model 3	83.11	**87.09**	84.80	81.43	86.98
Model 4	79.88	**82.94**	80.63	77.62	82.94
Model 5	81.04	**85.23**	84.48	78.59	84.77
Model 6	80.23	**85.86**	84.57	79.18	85.63
Model 7	80.38	**85.57**	83.69	77.74	84.62
Model 8	81.25	**85.28**	82.37	79.66	85.01
Model 9	82.43	**84.71**	82.96	80.21	84.31
Model 10	78.92	**82.89**	81.25	76.99	82.80
Model 11	79.98	**83.32**	81.96	77.36	83.29
Model 12	77.73	**81.29**	79.77	75.58	80.97

of the raw data and the features learned in the last FC layer of the proposed framework, as shown in Fig. 8. It can be seen that there is a serious problem of intra/inter-class distance change in such a large-scale fault diagnosis task due to the huge number of fault types.

To obtain the structure, other clustering methods can also be used, such as K-means. Therefore, the influence of the clustering methods is also discussed in this section. Concretely, a KM-PKT method is performed by using the K-means method and the PKT training scheme. The results are also shown in Table V. It can be seen that the performance of the method using K-means is slightly lower than the one used in the proposed method. The reason may be that K-means method is not been very stable and thus is not so well in dealing with high-dimensional data.

As a result, different fault states are heavily overlapped and can hardly be distinguished. While the fault features extracted by the proposed method can be easily distinguished or classified, the intra-class distances have been reduced and the inter-class distances have been enlarged, which verifies the superiority of the proposed framework in large-scale fault diagnosis task.

D. Comparison With Knowledge Transfer Methods

Knowledge transfer methods can generally be grouped into two parts: cross-domain transfer and intra-domain transfer, where the former extracts knowledge in a source domain

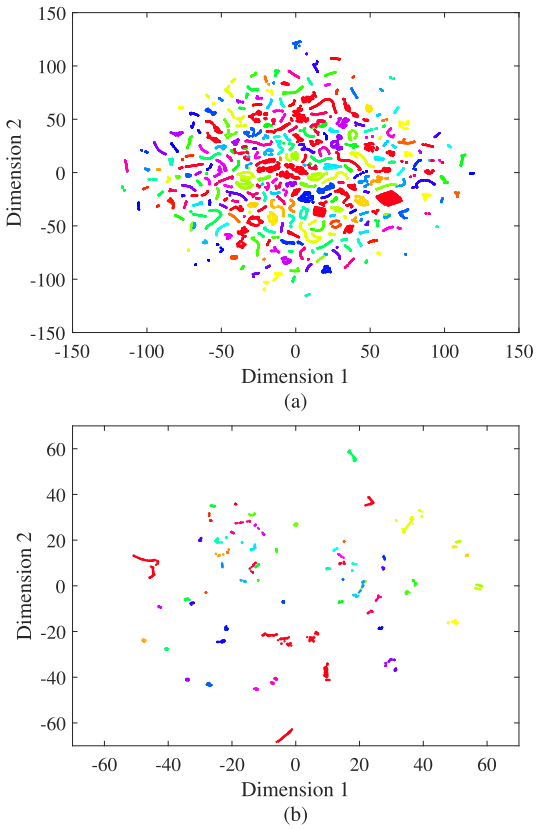


Fig. 8. Feature visualization via t-SNE for (a) original data space and (b) PKT-MCNN learning space with model 1.

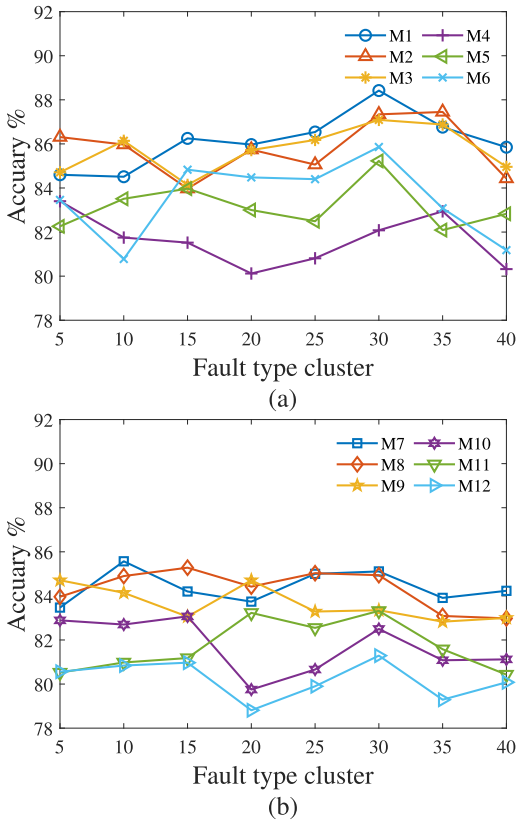


Fig. 9. Accuracies of PKT-MCNN under different fault type clusters with different CNN structures: (a) model 1–model 6 and (b) model 7–model 12.

and transfers it to the target domain, while the latter learns and transfers knowledge in the same domain. In our case, the model is supposed to extract the knowledge in the coarse-grained level and transfer such knowledge to the fine-grained level, where the two levels are in the same domain/data. Among many knowledge transfer methods, knowledge distillation (KD) is to transfer knowledge obtained by a teacher model to the student model is trained to mimic the prediction capabilities of the teacher. Here two KD models are performed. The vanilla KD is to train the student by minimizing the Kullback–Leibler (KL) divergence between the outputs of the teacher and the student, and AB-KD, which is proposed by Heo *et al.* [45], also minimizing the KL divergence between the logits of FC layer in addition to the vanilla KD.

The results on raw data and data with noise are shown in Tables III and IV, respectively. It can be seen that the proposed method can not only obtain the best performance, but also simpler to implement because KD methods need to train a larger model first and then proceed to transfer the knowledge to the small model. The reason why the proposed method can perform better may mainly be that coarse-grained knowledge can be learned by the model and is complementary to the target task, but such knowledge cannot be extracted by KD methods.

E. Analysis on the KT Controller

The proposed method designs a KT controller to effectively transfer the knowledge from the coarse to the fine. This can also be seen as weight the parameters of the loss terms. To this end, different weighting methods are also discussed via experiments on various CNN structures, including scaling weighting (scaling) [46] and uncertainty based self-learning weighting (self-learning) [40]. Specifically, we used the scaling weighting method to adjust both loss terms to the same scale. The uncertainty-based self-learning weighting method sets the parameters by calculating the uncertainty of two losses. The results are shown in Table VI. It can be seen that the proposed method is better than the two compared ones, which demonstrates the effectiveness of such a strategy.

TABLE VI
ACCURACY (%) COMPARISON BETWEEN DIFFERENT WEIGHTING STRATEGIES

# Model	Scaling	PKT	Self-learning
Model 1	82.47	**88.42**	84.47
Model 2	82.32	**87.45**	84.36
Model 3	83.11	**87.09**	84.98
Model 4	79.88	**82.94**	81.43
Model 5	81.04	**85.23**	84.22
Model 6	80.23	**85.86**	84.15
Model 7	80.38	**85.57**	83.91
Model 8	81.25	**85.28**	82.72
Model 9	82.43	**84.71**	81.69
Model 10	78.92	**82.89**	81.55
Model 11	79.98	**83.32**	82.35
Model 12	77.73	**81.29**	79.99

F. Ablation Study

To explore the influence of different stages of knowledge transfer, the ablation study was performed in this section

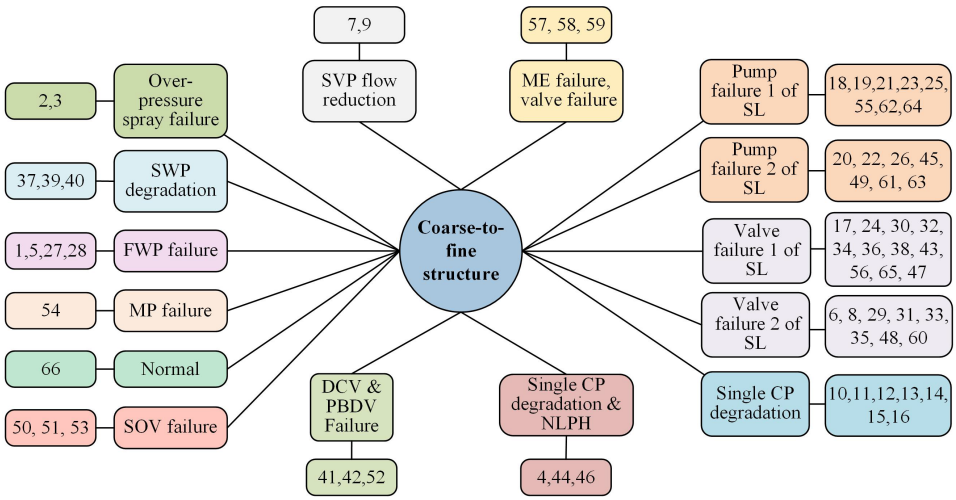


Fig. 10. Coarse-to-fine knowledge structure (SVP: speed variable pump; SWP: seawater pump; FWP: feed water pump; MP: main pump; ME: main engine; SOV: steam outer valve; CP: condensate pump; SL: secondary loop; NLPH: normal low-pressure heating; DCV: discharge cutoff valve; PBDV: pump bubble deaeration valve).

TABLE VII
KNOWLEDGE HIERARCHY WITH FIVE CLUSTERS

# Cluster	Coarse-grained fault concept	Fault type
1	Pump failure and Discharge cut-off valve failure	3, 4, 7, 9, 20, 21, 22, 25, 28, 40, 41, 42, 44, 45, 46, 52, 54, 61, 63
2	Single feed water pump failure	10, 11, 12, 14, 15, 16
3	Valve failure	17, 24, 26, 30, 32, 34, 36, 38, 43, 47, 49, 56, 65
4	Pump failure and valve failure	1, 2, 5, 6, 8, 18, 19, 23, 29, 31, 33, 35, 37, 39, 48, 55, 60, 62, 64
5	Steam outlet valve failure and Main engine quick closing valve failure	13, 27, 50, 51, 53, 57, 58, 59, 66

for all 12 PKT-MCNN methods. Concretely, there are three compared methods: 1) PKT-MCNN, which is trained by the proposed PKT algorithm; 2) CF-MCNN, which first trains the model on the coarse-grained task and then fine-tunes the fine-grained task; and 3) CM-MCNN, which first trains the model on the coarse-grained task and then simultaneously learns the coarse- and the fine-grained tasks via the designed PKT-MCNN model.

The results of the ablation study are shown in Table V. It can be seen that the performance of PKT-MCNN is significantly better than CF-MCNN and CM-MCNN, and CF-MCNN outperforms CM-MCNN by a large margin. Therefore, it can be concluded that the fine-grained training stage is indispensable for the fault diagnosis task. Moreover, CF-MCNN outperforms CM-MCNN by 2%–5%, and this may be because although the multitask learning stage is good to learn the two tasks simultaneously, it retains some features and discriminant information that are important for the coarse-grained task but not informative for the final fine-grained task. In addition, there is also an around 3% gap between PKT-MCNN and CF-MCNN, which empirically proves that the progressive multitask process controlled by the PKT can effectively transfer the useful coarse-grained knowledge to the fine-grained task.

G. Parameter Analysis

The number of coarse-grained fault concepts k is a hyper-parameter that may influence the performance of the proposed method. In this regard, the values of k were set to $\{5, 10, 15, 20, 25, 30, 35, 40\}$ among 12 PKT-MCNN methods to discuss the effect and optimal value of k. The diagnosis performance of the 12 PKT-MCNNs with different ks are shown in Fig. 9. It can be seen that the best k varies from the CNN architectures. Generally, the proposed method shows good performance when $k \leq 15$. This may be caused by the capacity of transferable coarse-grained knowledge to the fine-grained fault diagnosis task. When the value of k is small, for example, $k = 5$, the dissimilar fault types can also be clustered to a coarse node. Thus, the intra/inter class distance unbalance problem still exists in this node. In addition, the method may pay attention to more general and discriminant features that are less helpful for the fine-grained task, while too large values of k may provide too specific information, which makes the information of the coarse-grained task similar to that in the fine-grained task. Note that even though a relatively unreasonable k value would have an adverse effect on the training of the PKT-MCNN methods, their performances still exceed that of the conventional flat CNNs, which demonstrates the great advantage of the proposed approach for large-scale fault diagnosis task.

H. Analysis on the Extracted Knowledge Structure

Since the proposed approach is data-driven, the coarse-to-fine knowledge structure can be extracted without human intervention. To interpret the rationale behind the learning mechanisms, the automatically learned knowledge is analyzed in this section. The knowledge structure with $k = 15$ is illustrated in Fig. 10, and the numbers in the figure are the ordinal of fault types. It can be seen that the coarse-grained fault concepts are extracted based on two main evidences: fault types from the same components and fault types with similar operational characteristics. For example, in Fig. 10, SVP, SWP, MP, and CP are different components of the nuclear power

TABLE VIII
KNOWLEDGE HIERARCHY WITH 10 CLUSTERS

# Cluster	Coarse-grained fault concept	Fault type
1	Steam outlet valve failure and normal status	50, 51, 53, 66
2	Valve failure 1	17, 24, 30, 32, 34, 36, 38, 43, 47, 56, 65
3	Valve failure 2	1, 2, 5, 6, 8, 29, 31, 33, 35, 39, 48
4	Main engine quick closing valve failure	57, 58, 59
5	Speed variable pump failure	7, 9, 54
6	Pump failure	20, 22, 26, 40, 45, 49, 61, 63
7	Single feed water pump failure	10, 11, 12, 13, 14, 15, 16
8	Condensate pump and circulating pump failure	18, 19, 21, 22, 23, 37, 55, 60, 62, 64
9	Discharge cut-off valve failure	28, 41, 42, 52
10	Feed water and condensate pump failure	3, 4, 25, 27, 44, 46

TABLE IX
KNOWLEDGE HIERARCHY WITH 20 CLUSTERS

# Cluster	Coarse-grained fault concept	Fault type
1	Speed variable pump failure	7, 9
2	Valve failure 1	17, 24, 30, 32, 34, 36, 38, 43, 56, 65
3	Valve failure 2	6, 8, 29, 31, 33, 35, 48, 60
4	Steam outlet valve failure	50, 51, 53
5	Main engine quick closing valve failure	57, 58, 59
6	Seawater pump failure	37, 39, 40
7	Over-pressure spray failure	2, 3
8	Single feed water pump failure	10, 11, 15
9	Single feed water pump failure	12, 13
10	Single feed water pump failure	14, 16
11	Main pump failure	54
12	Condensate pump failure	44, 46
13	Pump failure 1	20, 22, 45, 49, 61, 63
14	Pump failure 2	18, 19, 21, 23, 55, 64
15	Discharge cut-off valve failure	41, 42, 52
16	Multiple feed water pump failure	27, 28
17	Normal low-pressure heating and feed water pump failure	4, 25
18	Multiple main circulating pump failure	62
19	Feed water pump failure	1, 5, 26, 47
20	Normal status	66

system, and they were extracted as mutually exclusive superordinate fault concepts. By contrast, faults from pump fault and valve fault of SL were assigned to two different superordinate fault concepts, which implies that some faults are very different even if they come from the same component. It is worth noting that the component information plays a major role of knowledge extraction, because fault types with similar operational characteristics also occur on the same component.

Although results vary from different clusters and values of k, there are some common rules for them. First, fault types from the same components are often assigned to the same groups. For example, there are two groups in all the coarse-to-fine network structures that follow such a rule: 1) 17, 24, 30, 32, 34, 36, 43, 56, 65; and 2) 20, 22, 45, 61, 63. The numbers here are the ordinal of fault types. It can be seen that all the fault types in group 1) are valve failures and those in group 2) are pump failures. Second, some fault types are split into smaller fault concepts with the increase of the cluster number. This can be commonly observed in most results from Tables VII–IX. The rationale of this phenomenon is that the proposed model clusters the fault types based on the similarity between them, and some clusters of the structures with more coarse-grained nodes (e.g., $k = 20$) are the same with the subclusters of those with fewer coarse-grained nodes (e.g., $k = 5$).

V. CONCLUSION

In this article, a novel coarse-to-fine knowledge transfer framework is proposed for large-scale fault diagnosis. First, the coarse-to-fine structure is learned adaptively by the proposed structure learning algorithm. Then, to integrate the processes of learning each task and transferring the useful information from the coarse-grained to the fine-grained, a multitask CNN is designed by sharing the representations and owing different classifiers to each task. Third, a PKT algorithm is designed to transfer the discriminant information and turn the attention of the PKT-MCNN from the coarse-grained to the fine-grained fault diagnosis task progressively. Finally, a large-scale dataset of a nuclear power system is collected and analyzed by the proposed approach. The experimental results illustrated the effectiveness and superiority of the proposed method, which not only intelligently learned a reasonable knowledge structure that coincides with the physical composition of the nuclear power system, but also has great advantages on dealing with the special physical background of large-scale fault diagnosis via extracting and transferring the coarse-grained knowledge to the final fine-grained diagnosis task, and thus can be a promising tool in future research of industrial big data analysis.

REFERENCES

[1] N. Qin, K. Liang, D. Huang, L. Ma, and A. H. Kemp, "Multiple convolutional recurrent neural networks for fault identification and performance degradation evaluation of high-speed train bogie," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5363–5376, Dec. 2020.

[2] R. Liu, F. Wang, B. Yang, and S. J. Qin, "Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3797–3806, Jun. 2020.

[3] K. Zhong, M. Han, T. Qiu, and B. Han, "Fault diagnosis of complex processes using sparse kernel local Fisher discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1581–1591, May 2019.

[4] L. Yao, W. Shao, and Z. Ge, "Hierarchical quality monitoring for large-scale industrial plants with big process data," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 31, 2020, doi: 10.1109/TNNLS.2019.2958184.

[5] Z. Liu and L. Zhang, "A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings," *Measurement*, vol. 149, Jan. 2020, Art. no. 107002.

[6] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021.

[7] Q. Jiang, S. Yan, H. Cheng, and X. Yan, "Local-global modeling and distributed computing framework for nonlinear plant-wide process monitoring with industrial big data," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 21, 2020, doi: 10.1109/TNNLS.2020.2985223.

[8] M. Zhao, S. Zhong, X. Fu, B. Tang, S. Dong, and M. Pecht, "Deep residual networks with adaptively parametric rectifier linear units for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 68, no. 3, pp. 2587–2597, Mar. 2021.

[9] J. Jiao, M. Zhao, J. Lin, and C. Ding, "Deep coupled dense convolutional network with complementary data for intelligent fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9858–9867, Dec. 2019.

[10] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[11] D. Erhan, P. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," *J. Mach. Learn. Res.*, vol. 5, pp. 153–160, Apr. 2009.

[12] R. Babbar, I. Partalas, E. Gaussier, and M. R. Amini, "On flat versus hierarchical classification in large-scale taxonomies," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1824–1832.

[13] T. Zhao *et al.*, "Embedding visual hierarchy with deep networks for large-scale visual recognition," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4740–4755, Oct. 2018.

[14] X. Chu *et al.*, "Distance metric learning with joint representation diversification," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, vol. 1, 2020, pp. 1962–1973.

[15] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[16] S. Xing, Y. Lei, S. Wang, and F. Jia, "Distribution-invariant deep belief network for intelligent fault diagnosis of machines under new working conditions," *IEEE Trans. Ind. Electron.*, vol. 68, no. 3, pp. 2617–2625, Mar. 2021.

[17] H. Wang, M.-J. Peng, J. Wesley Hines, G.-Y. Zheng, Y.-K. Liu, and B. R. Upadhyaya, "A hybrid fault diagnosis methodology with support vector machine and improved particle swarm optimization for nuclear power plants," *ISA Trans.*, vol. 95, pp. 358–371, Dec. 2019.

[18] X. Xu, D. Cao, Y. Zhou, and J. Gao, "Application of neural network algorithm in fault diagnosis of mechanical intelligence," *Mech. Syst. Signal Process.*, vol. 141, Jul. 2020, Art. no. 106625.

[19] P. Liu, J. Wang, and Z. Guo, "Multiple and complete stability of recurrent neural networks with sinusoidal activation function," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 229–240, Jan. 2021.

[20] Z. An, S. Li, J. Wang, and X. Jiang, "A novel bearing intelligent fault diagnosis framework under time-varying working conditions using recurrent neural network," *ISA Trans.*, vol. 100, pp. 155–170, May 2020.

[21] H. Liu, J. Zhou, Y. Zheng, W. Jiang, and Y. Zhang, "Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders," *ISA Trans.*, vol. 77, pp. 167–178, Jun. 2018.

[22] L. Li, J. Yan, H. Wang, and Y. Jin, "Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1177–1191, Mar. 2021.

[23] 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.

[24] Y. Wang, Z. Pan, X. Yuan, C. Yang, and W. Gui, "A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network," *ISA Trans.*, vol. 96, pp. 457–467, Jan. 2020.

[25] C. Shen, J. Xie, D. Wang, X. Jiang, J. Shi, and Z. Zhu, "Improved hierarchical adaptive deep belief network for bearing fault diagnosis," *Appl. Sci.*, vol. 9, no. 16, p. 3374, Aug. 2019.

[26] S. K. Khare and V. Bajaj, "Time-frequency representation and convolutional neural network-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021.

[27] R. Liu, B. Yang, and A. G. Hauptmann, "Simultaneous bearing fault recognition and remaining useful life prediction using joint-loss convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 87–96, Jan. 2020.

[28] S. Guo, B. Zhang, T. Yang, D. Lyu, and W. Gao, "Multitask convolutional neural network with information fusion for bearing fault diagnosis and localization," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 8005–8015, Sep. 2020.

[29] Z. Chen, K. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 339–349, Jan. 2020.

[30] F. Wang, R. Liu, Q. Hu, and X. Chen, "Cascade convolutional neural network with progressive optimization for motor fault diagnosis under nonstationary conditions," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2511–2521, Apr. 2021.

[31] M. Azamfar, J. Singh, I. Bravo-Imaz, and J. Lee, "Multisensor data fusion for gearbox fault diagnosis using 2-D convolutional neural network and motor current signature analysis," *Mech. Syst. Signal Process.*, vol. 144, Oct. 2020, Art. no. 106861.

[32] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei, "Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3450–3457.

[33] Y. Wang *et al.*, "Deep fuzzy tree for large-scale hierarchical visual classification," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1395–1406, Jul. 2020.

[34] A. Li, T. Luo, Z. Lu, T. Xiang, and L. Wang, "Large-scale few-shot learning: Knowledge transfer with class hierarchy," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7212–7220.

[35] X. Wei, C. Zhang, L. Liu, C. Shen, and J. Wu, "Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 575–591.

[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[37] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 527–538.

[38] D. Lin *et al.*, "ZigZagNet: Fusing top-down and bottom-up context for object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7490–7499.

[39] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1666–1674.

[40] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.

[41] H. Wu and J. Zhao, "Deep convolutional neural network model based chemical process fault diagnosis," *Comput. Chem. Eng.*, vol. 115, no. 12, pp. 185–197, Jul. 2018.

[42] A. Klausen and K. G. Robbersmyr, "Cross-correlation of whitened vibration signals for low-speed bearing diagnostics," *Mech. Syst. Signal Process.*, vol. 118, pp. 226–244, Mar. 2019.

[43] W. Yu and C. Zhao, "Broad convolutional neural network based industrial process fault diagnosis with incremental learning capability," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5081–5091, Jun. 2020.

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

[45] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, vol. 33, no. 1, pp. 3779–3787.

[46] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 793–802.

Yu Wang received the B.S. degree in communication engineering, the M.S. degree in software engineering, and the Ph.D. degree in computer applications and techniques from Tianjin University, Tianjin, China, in 2013, 2016, and 2020, respectively.

He is currently an Assistant Professor with Tianjin University and was an Outstanding Visitor Scholar of University of Waterloo, Waterloo, ON, Canada, in 2019. His research interests focus on hierarchical learning and large-scale classification in industrial scenarios and computer vision applications, data mining, and machine learning. He has published many peer-reviewed papers in world-class conferences and journals, such as IEEE TRANSACTIONS ON FUZZY SYSTEMS (TFS), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), and so on.

Ruonan Liu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2013, 2015, and 2019, respectively.

She was a Post-Doctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, in 2019. She is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Her research interests include machine learning, intelligent manufacturing, and computer vision.

Di Lin (Member, IEEE) received the bachelor's degree in software engineering from Sun Yat-sen University, Guangzhou, China, in 2012, and the Ph.D. degree from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2016.

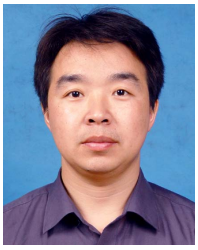
He is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests are computer vision and machine learning.

Dongyue Chen received the B.S. degree in transportation equipment information engineering and the M.S. degree in precision instruments and machinery from Southwest Jiaotong University, Chengdu, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree in computer applications and techniques with the College of Intelligence and Computing, Tianjin University, Tianjin, China.

Her research interests include deep learning, condition monitoring, and fault diagnosis of mechanical system.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2013.

He is currently a Research Assistant Professor with the Hong Kong Polytechnic University, Kowloon, Hong Kong. His current research interests include image/video stylization, GPU acceleration, and creative media. He has one image/video processing national invention patent and has excellent research project reported worldwide by *ACM TechNews*.



Qinghua Hu (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively.

He was a Post-Doctoral Fellow with the Department of Computing, Hong Kong Polytechnic University, from 2009 to 2011. He is currently the Dean of the School of Artificial Intelligence, the Vice Chairman of the Tianjin Branch of China Computer Federation, the Vice Director of the SIG Granular Computing and Knowledge Discovery, and the Chinese Association of Artificial Intelligence. He is currently supported by the Key Program, National Natural Science Foundation of China. He has published over 200 peer-reviewed papers. His current research is focused on uncertainty modeling in big data, machine learning with multimodality data, intelligent unmanned systems.

He is an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Acta Automatica Sinica*, and *Energies*.

C. L. Philip Chen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988.

He is currently a Chair Professor and the Dean of the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Being a Program Evaluator of the Accreditation Board of Engineering and Technology Education (ABET) in the U.S., for computer engineering, electrical engineering, and software engineering programs, he successfully architects the University of Macau's Engineering and Computer Science programs receiving accreditations from Washington/Seoul Accord through Hong Kong Institute of Engineers (HKIE), of which is considered as his utmost contribution in engineering/computer science education for Macau as the former Dean of the Faculty of Science and Technology. His current research interests include systems, cybernetics, and computational intelligence.

Dr. Chen was a recipient of the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University (in 1988), after he graduated from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA, in 1985. He received IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learning. He is also a highly cited researcher by Clarivate Analytics in 2018 and 2019. He is a Fellow of AAAS, IAPR, CAA, and HKIE, and a member of Academia Europaea (AE), European Academy of Sciences and Arts (EASA), and International Academy of Systems and Cybernetics Science (IASCYS). He was the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013, the Editor-in-Chief of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS (2014–2019), and currently, he is the Editor-in-Chief of the IEEE TRANSACTIONS ON CYBERNETICS, and an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS. He was the Chair of TC 9.1 Economic and Business Systems of International Federation of Automatic Control from 2015 to 2017, and currently is a Vice President of Chinese Association of Automation (CAA).