Adversarial Multimodal Contrastive Learning for Robust Industrial Fault Diagnosis

Rongyao Cai, Kexin Zhang, *Member, IEEE*, Hanchen Tai, Yang Zhou, Yuanyuan Ding, Chunlin Zhou, and Yong Liu, *Member, IEEE*

***Abstract*—Fault diagnosis (FD) techniques leveraging self-supervised contrastive learning (SSCL) have demonstrated significant potential in industrial scenarios due to their reduced dependence on manually annotated data. However, the existing SSCL algorithms primarily focus on establishing complex similarity relationships among unimodal augmented views. These unimodal SSCL approaches are particularly vulnerable to learning shallow, domain-dependent spurious features in the training data rather than more intrinsic and essential features. Consequently, such spurious features may cause the algorithm failure when encountering distribution shift issues resulting from environmental perturbations or changes in working conditions. To address this challenge, we propose adversarial multimodal contrastive learning (AMMCL), a novel approach designed to extract robust and generalizable multimodal representations from time series and their corresponding spectrograms. AMMCL utilizes intermodal contrastive learning and adversarial training strategy to align modal-invariant features from both elementwise and setwise perspectives. These essential features are beneficial for intradomain and cross-domain FD tasks. Furthermore, a slice segmentation processing (SSP) method based on dominant frequency is employed to enhance model's ability to recognize varying patterns within time series. AMMCL is first evaluated on intradomain and cross-domain FD tasks using the Gearbox and XJTU-SY datasets, where it outperforms nine existing FD algorithms in terms of performance. Additionally, AMMCL is compared with ten other valve stiction detection algorithms on International Stiction Database (ISDB) dataset, successfully identifying the most loop states (23 out of 26). Finally, the trained AMMCL model on the ISDB dataset is implemented in actual industrial valve detection, demonstrating the feasibility and practicality of AMMCL in real industrial scenarios.**

***Index Terms*—Adversarial training, cross-domain experiments, fault diagnosis (FD), multimodal contrastive learning, time series.**

Received 13 December 2024; revised 7 July 2025; accepted 22 August 2025. Date of publication 18 September 2025; date of current version 30 September 2025. This work was supported in part by the Key Research and Development Project of Zhejiang Province under Grant 2024C01172 and in part by the Young Scientists Fund of National Natural Science Foundation of China under Grant 62503419. The Associate Editor coordinating the review process was Dr. Zhuyun Chen. *(Corresponding authors: Kexin Zhang; Chunlin Zhou.)*

Rongyao Cai, Hanchen Tai, Yuanyuan Ding, Chunlin Zhou, and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China (e-mail: rycai@zju.edu.cn; 22232108@zju.edu.cn; 22232062@zju.edu.cn; c_zhou@zju.edu.cn; yongliu@iipc.zju.edu.cn).

Kexin Zhang is with Huzhou Institute of Zhejiang University, Huzhou 313000, China, and also with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China (e-mail: zhangkexin@zju.edu.cn).

Yang Zhou is with the School of Automation, China University of Geosciences, Wuhan 430074, China (e-mail: zhouyang@cug.edu.cn).

Digital Object Identifier 10.1109/TIM.2025.3608323

I. Introduction

FAULT diagnosis (FD) plays a pivotal role in ensuring the stability of industrial processes. Its primary purpose is to identify the equipment's operational states, including system states and fault types, through time-series data from monitoring sensors. In recent years, deep learning-based FD algorithms have gained popularity in various intelligent systems due to their adaptability [1], [2], [3], [4], [5], [6]. These data-driven algorithms aim to obtain feature extractors that suit the specific demands of diverse downstream tasks. However, developing reliable deep learning models often requires substantial annotated data, which may be constrained by factors, such as domain expertise, increasing labor costs, and data privacy concerns. To tackle this challenge, contrastive learning has emerged as a promising approach, demonstrating low dependence on labeled data.

The self-supervised contrastive learning (SSCL) framework, comprising pretraining and fine-tuning stages, has rapidly gained traction in computer vision and natural language processing. This paradigm is appealing as it can mitigate the dependency for annotated data, enhance model generalization capability, and adapt to various tasks. Thus, incorporating this learning paradigm into FD models is a viable strategy. Current SSCL-based FD research primarily focuses on developing intricate similarity architectures [7], [8], [9], [10], [11] and mining (MIN) multigranularity representations [12], [13], [14]. For example, Zhang et al. [2] proposed an expert-knowledge-guided novel framework to address the task-agnostic issue in SSCL. Cai et al. [6] introduced a Gaussian distance metric in SSCL to measure the radial distribution, aiding cosine similarity, and experimented on a diverse industrial dataset. Peng et al. [11] used a linear classifier to identify chemical processes faults after obtaining the informative SSCL representations. These studies have harnessed contrastive learning's potential, achieving promising results in FD tasks.

While SSCL methods achieve optimal performance in FD tasks, challenges remain, particularly in pointwise processing and unimodal feature extraction. For FD tasks, the pattern changes in signals caused by faults are crucial for identifying fault types and tracing their occurrence. However, pointwise processing treats individual time-series points as model's basic input unit. Single points lack statistical significance, potentially causing the model to focus excessively on local changes and insufficiently recognize global patterns. Meanwhile, timestamp misalignment across different channels can lead to incorrect interchannel relationship MIN under pointwise processing. SSCL methods extract features by comparing the similarity among positive and negative pairs. Typically, augmented views from the same sample are positive, while the others are negative. When models are trained on unimodal augmented pairs, they predominantly focus on surface-level features. This limitation is closely tied to the critical role of augmentation techniques in determining the performance of unimodal methods [15]. Although more aggressive augmentation strategies might boost the performance, they often disrupt the inherent seasonal or trend patterns fundamental to time-series analysis. For instance, cyclical behaviors and long-term trends commonly found in time-series data are highly vulnerable to distortion by strong augmentations. Furthermore, common augmentation methods primarily modify the surface-level shape of sequences. These transformations do not engage with the underlying semantic patterns that neural networks need to discern.

To address the above issues, we proposed a novel adversarial multimodal contrastive learning (AMMCL) framework for FD tasks in intelligent systems monitoring. The AMMCL framework extracts the modal features from time-series data and corresponding spectrogram images. Time series and spectrograms provide distinct yet semantically equivalent descriptions of the same object. To refine and integrate essential information between modal features for FD tasks, intermodal contrastive learning and an adversarial training strategy are employed. This approach enhances the modal-invariant part and inhibit modal-specific part from both elementwise and setwise perspective. The modal-invariant information is intrinsic and shared across modalities, contributing to understanding the operational state of systems. In contrast, modal-specific information refers to the distinct features inherent to a particular modality, including temporal locality and phase relationships in time-domain signals and harmonic structures in frequency-domain signals. Finally, we introduce slice segmentation processing (SSP) based on dominant frequency for time series, treating each slice as the basic processing unit of the model. This approach enhances the model's ability to recognize signal patterns more effectively and mitigate the misalignment issue between channels. The main contributions are as follows.

1) A novel AMMCL framework is proposed for industrial FD tasks. AMMCL extracts robust and fundamental cross-modal features from time series and spectrograms, adapting better to dynamic industrial environments.
2) An adversarial training strategy is employed to enhance the modal-invariant features, which represent the essential information shared across modalities, while suppressing the modal-specific parts, which refer to the unique characteristics of each individual modality.
3) SSP based on the dominant frequency term is utilized, considering slices as the basic input units of models to better mine pattern changes in time series.
4) Extensive experiments on three public datasets indicate that AMMCL extracts more robust and essential features for FD tasks. Beyond that, we deploy the trained AMMCL in real control systems to evaluate its practicality for industrial deployment.

The remainder of this article is structured as follows. Section II introduces the SSP technique, multimodal contrastive learning, and adversarial training strategy. Section III details the comprehensive experiments conducted to evaluate the performance of AMMCL. Finally, Section IV summarizes our works and concludes this article.

II. Methodologies

A. Problem Definition

Consider $X_i = \{x_1, x_2, \ldots, x_D\} \in \mathbb{R}^{L \times D}, i = 1, 2, \ldots, N$ as a single multivariate time-series sample with length L and dimension D, where $y_i \in \mathbb{R}$ is the ground truth for sample X_i. The dataset $\mathcal{X} = \mathcal{X}^{an} \cup \mathcal{X}^{un}$ comprises N samples, including normal and abnormal ones. Here, $\mathcal{X}^{an}$ and $\mathcal{X}^{un}$ denote the subsets of annotated and unannotated samples with N^{an} and N^{un} samples, respectively, and $N^{an} \ll N^{un}$. Our goal is to develop a feature extraction framework for multimodal data that can be utilized for industrial FD tasks. This framework should eliminate the need of manual feature engineering and establish a data-driven FD model based on $\mathcal{X}$.

B. Overall Framework

The pretraining architecture of the proposed AMMCL is depicted in Fig. 1. First, the original time series X_i is converted into spectrograms I_i via the short-time Fourier transform (STFT) algorithm. Then, augmented views X_i^s and X_i^w are generated from X_i using time-series augmentation techniques, such as jittering and permutation. Similarly, I_i^s and I_i^w are created from I_i through image augmentation methods, such as noising and masking. Features f_i and z_i are extracted from these augmented views using dedicated time series and spectrogram feature extractors, i.e., FE_X and FE_I, where are central to the pretraining stage and have their weights frozen during fine-tuning, as shown in Fig. 2. Intramodal InfoNCE loss functions $\mathcal{L}_X$ and $\mathcal{L}_I$ are applied to regulate similarity within each modality.

To extract high-level, generalizable features, an intermodal InfoNCE loss $\mathcal{L}_{XI}$ and an adversarial training strategy $\mathcal{L}_D$ are employed to align cross-modal features from both elementwise and setwise perspectives. $\mathcal{L}_{XI}$ treats f_i and its corresponding z_i as a positive pair, drawing them closer. $\mathcal{L}_D$ distinguishes the modality of features, reducing the distribution margin between f_i and z_i.

In the fine-tuning stage (see Fig. 2), the fusion feature $\hat{f}_i$ is generated by merging f_i and z_i with a fusion factor α. Downstream classifiers are then activated to predict the label $\hat{y}_i$ of X_i based on $\hat{f}_i$. Specifically, FE_X and FE_I are frozen and excluded from backpropagation, while the downstream classifiers are trained to minimize the cross-entropy loss.

C. Slice Segmentation Processing

While pointwise processing approaches extract information by treating individual time steps as fundamental units [16], [17], this methodology has inherent limitations for FD tasks. Effective FD requires comprehensively identifying pattern compositions within industrial signals to enable accurate fault

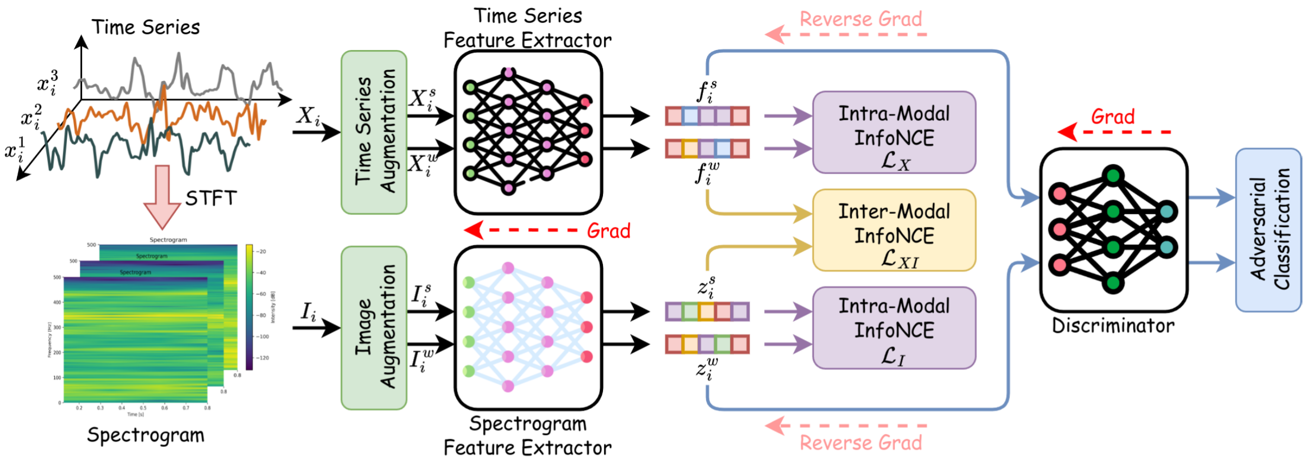

Time Series
x_i^3
x_i^2
x_i^1
STFT
Spectrogram
X_i
I_i
Time Series Augmentation
Image Augmentation
X_i^s
X_i^w
I_i^s
I_i^w
Time Series Feature Extractor
Spectrogram Feature Extractor
Grad
f_i^s
f_i^w
z_i^s
z_i^w
Reverse Grad
Intra-Modal InfoNCE $\mathcal{L}_X$
Inter-Modal InfoNCE $\mathcal{L}_{XI}$
Intra-Modal InfoNCE $\mathcal{L}_I$
Grad
Discriminator
Adversarial Classification

Fig. 1. Pretraining stage of AMMCL. All modules are activated for backpropagation.

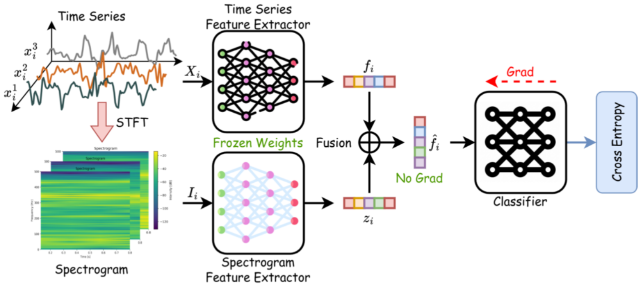


Fig. 2. Fine-tuning stage of AMMCL. The time series and spectrogram feature extractors are frozen, and the downstream classifiers are activated for backpropagation.

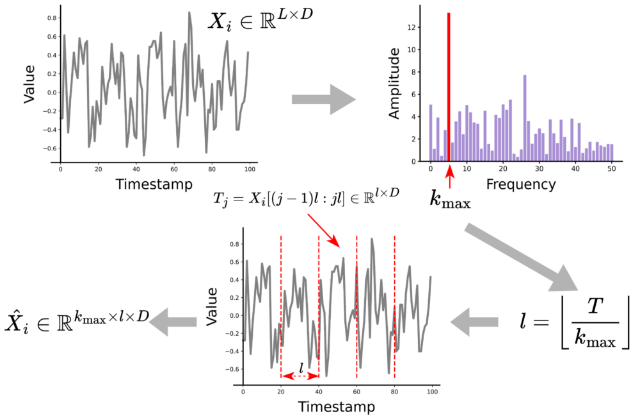

$X_i \in \mathbb{R}^{L\times D}$
Value
Timestamp
Amplitude
Frequency
$k_{\max}$
$T_j = X_i[(j-1)l : jl] \in \mathbb{R}^{l\times D}$
$l = \left\lfloor \frac{T}{k_{\max}} \right\rfloor$
l
Value
Timestamp
$\hat{X}_i \in \mathbb{R}^{k_{\max}\times l\times D}$

Fig. 3. Operation of the proposed SSP method based on dominant frequency.

classification and traceability. The critical limitation of pointwise analysis is statistical insufficiency of isolated temporal points, which restricts the model's receptive field. Such localized processing fails to capture essential global characteristics of time-series data, including critical trend variations and periodic patterns vital for robust fault detection [18].

Furthermore, time misalignment across sensor channels often occurs due to sampling errors or human errors during data collection. When processed using pointwise methods, this misaligned data force measurements from different timestamps to be merged into a single data token. Mismatched bundling causes the model to incorrectly associate sensor readings that are not temporally aligned, leading to the learning of spurious interchannel relationships contradictory to the actual operational behavior of industrial equipment.

To overcome the limitations, we propose the SSP method, which treats temporal slices of length l as fundamental processing units rather than individual timestamps. This approach ensures statistically meaningful input units while enhancing robustness against temporal misalignment across sensor channels. However, arbitrary sequence segmentation risks disrupting long-term periodic patterns, constraining the model to short-term features. To preserve critical cyclical information, we develop an adaptive segmentation strategy based on fast Fourier transform (FFT) analysis. As shown in Fig. 3, our method first identifies the dominant frequency term $k_{\max}$ through spectral decomposition of the original series X_i and then determines the optimal slice length l corresponding to the principal frequency. The SSP operation is defined as follows:

$$H(k) = \sum_{n=0}^{L-1} X[n]\, e^{-j2\pi kn/L} \tag{1}$$

$$k_{\max} = \arg\max_{k} |H(k)| \tag{2}$$

$$l = \left\lfloor \frac{L}{k_{\max}} \right\rfloor \tag{3}$$

where $H(k)$ is the kth term of Fourier series, $|\cdot|$ is the modulo operation, and $\lfloor\cdot\rfloor$ is the downward rounding arithmetic.

After data augmentation, we implement the SSP operation by participating each augmented series $X_i \in \mathbb{R}^{L\times D}$ into temporal slices T_i of length l. These slices are then temporally stacked to form a processed 3-D tensor $\hat{X}_i \in \mathbb{R}^{K_{\max}\times l\times D}$

$$\hat{X}_i = \text{Stack}\left(T_1, T_2, \ldots, T_{k_{\max}}\right) \in \mathbb{R}^{k_{\max}\times l\times D} \tag{4}$$

where Stack$(\cdot)$ denotes the stacking function in dim = 0.

D. Multimodal Contrastive Learning

Traditional contrastive learning methods usually compare positive and negative pairs within the same modality. While

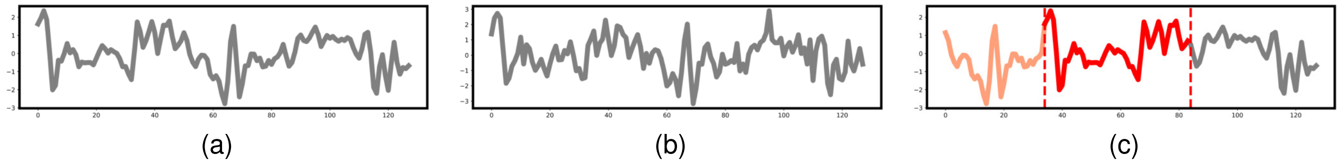


Fig. 4. Time-series augmentation techniques. (a) Original time series. (b) Time series after weak augmentation (jitter). (c) Time series after strong augmentation (permutation).

stronger augmentation techniques can boost feature extraction efficiency for challenging tasks [15], they are predominantly used in intramodal comparisons. Models trained in this manner aim to identify features that are invariant to the applied augmentations. However, these features often lack the deeper semantic meaning necessary for a comprehensive and context-aware understanding. Such shallow features can lead to algorithm failure when encountering distribution shift problems caused by environmental perturbations. Furthermore, overusing high-perturbation augmentation methods may distort extracted features due to distributional divergence.

We choose spectrograms as a distinct modality from time series because they can visually and dynamically decompose the frequency components of time series via the STFT algorithm, offering valuable insights for patterns identification. Spectrograms and time series share equivalent temporal semantics but in different forms. Inspired by multimodal image–text transformation [19], [20], applying SSCL between spectrogram and time series can yield meaningful cross-modal representations that capture data's inherent structure and characteristics. Spectrogram provides a global view of the system's dynamic components, while time series captures local changes more effectively.

Time series exhibits strict temporal dependencies and local continuity. Jitter and permutation are the common augmentation techniques for time series, as shown in Fig. 4. Jitter adds Gaussian noise with mean $\mu = 0$ and standard deviation σ_1 to the original data, enhancing the robustness of model to noise and minor variations. Permutation randomly rearranges n segments X_i^j of the time series X_i, guiding the model to recognize patterns irrespective of their specific order

$$X_i^w = X_i + \text{Norm}(\mu, \sigma_1) \tag{5}$$

$$X_i = \left[X_i^1, X_i^2, \ldots, X_i^n\right] \tag{6}$$

$$X_i^s = \text{Perm}\left(X_i^1, X_i^2, \ldots, X_i^n\right) \tag{7}$$

where n denotes the number of segments, Norm$(\cdot,\cdot)$ represents the Gaussian noise, and Perm$(\cdot)$ is the out-of-order function.

Spectrograms, as 2-D images, leverage spatial redundancy and hierarchical semantics. In vision contrastive [21], pixel noise and random masking are the standard techniques. Pixel noise is added by introducing Gaussian noise with mean $\mu = 0$ and standard deviation σ_2 to the original figure, mimicking real-world imaging artifacts (e.g., low-light noise). Random masking covers pixels with a probability of p, forcing the model to learn from partial contexts and improving robustness to occlusions

$$I_i^w = I_i + \text{Norm}(\mu, \sigma_2) \tag{8}$$

$$I_i^s = I_i \odot \text{Bool}(p) \tag{9}$$

where Bool$(\cdot)$ is the Boolean mask matrix and $\odot$ denotes the Hadamard product.

InfoNCE losses $\mathcal{L}_X$ and $\mathcal{L}_I$ are employed to constrain the similarity of intramodal positive pairs. Intramodal contrastive learning enhances the utilization of intramodal information and simplifies cross-modal contrastive learning

$$\mathcal{L}_X = \mathbb{E}\left[-\log \frac{\exp\left(\text{sim}\left(f_i^s, f_i^w\right)/\tau\right)}{\sum_{i \neq j} \exp\left(\text{sim}\left(f_i, f_j\right)/\tau\right)}\right] \tag{10}$$

$$\mathcal{L}_I = \mathbb{E}\left[-\log \frac{\exp\left(\text{sim}\left(z_i^s, z_i^w\right)/\tau\right)}{\sum_{i \neq j} \exp\left(\text{sim}\left(z_i, z_j\right)/\tau\right)}\right] \tag{11}$$

where τ is the temperature coefficient to regulate the convexity.

The features f_i and z_i though from different modalities share the same semantics and are refined from specific time series and their corresponding spectrogram, respectively. To constrain cross-modal positive pair f_i and z_i, a novel $\mathcal{L}_{XI}$ is proposed as follows:

$$\mathcal{L}_{XI} = \mathbb{E}\left[-\log \frac{\exp\left(\text{sim}\left(f_i, z_i\right)/\tau\right)}{\sum_{i \neq j} \exp\left(\text{sim}\left(f_i, z_j\right)/\tau\right)}\right]. \tag{12}$$

E. Adversarial Training Strategy

Cross-modal positive pairs, though semantically consistent, contain modal-specific components that distinguish them from modal-invariant parts. Modal-specific components refer to features that are intrinsic to a particular modality. For instance, time-domain signals are characterized by temporal locality and phase relationships that determine the timing and alignment of signal components. Meanwhile, frequency-domain signals are often distinguished by harmonic structures, which provide details about the periodic components of the signal, including their frequencies, amplitudes, and phases. These features are fundamental to how each modality represents the underlying data and phenomena.

Let $\mathcal{M} = \{M_X, M_I\}$ denote a set of modalities, where M_X and M_I are linked to specific feature spaces $\mathcal{F}_X$ and $\mathcal{F}_I$. Modal-invariant features are conceptualized as a shared latent representation $\mathcal{F}_{\text{inv}}$, capturing the common semantic information across all modalities. Mathematically, we aim to adapt feature extractors FE$_X$ and FE$_I$ that map the input data X and I into a unified feature space $\mathcal{F}_{\text{inv}}$, such that

$$\mathcal{F}_{\text{inv}} = \text{FE}_X(X) \cap \text{FE}_I(I). \tag{13}$$

In adversarial training, the goal is to minimize the distributional discrepancies of modal-specific features in the shared latent space. Specifically, an adversarial discriminator $D(\cdot)$

is employed to distinguish the origin of features, i.e., to identify the modality from which they are derived. The process is formulated as a minimization optimization problem $\mathcal{L}_D$, incorporating a gradient reversal layer $R(\cdot)$

$$R(x) = x, \quad \frac{\partial R}{\partial x} = -I \tag{14}$$

$$\min \mathcal{L}_D = \min_{\theta} \mathop{\mathbb{E}}_{h \in \{f,z\}} \left[-\sum c_i \log (D(R(h_i))) \right] \tag{15}$$

where h generalizes f and z and θ represents the model parameters. Specifically, the pseudo label c_i is 1 if h_i corresponds to the feature f_i from modality M_X and 0 if h_i corresponds to the feature z_i from modality M_I. Our objective is to minimize the classification loss, ensuring the alignment of feature distributions across modalities and promoting the learning of modal-invariant features.

The synergy of $\mathcal{L}_{XI}$ and $\mathcal{L}_D$ within the AMMCL framework ensures effective features alignment across modalities without performance degradation after fusion. $\mathcal{L}_{XI}$ strongly constrains cross-modal positive pairs by pulling them together elementwise. In contrast, $\mathcal{L}_D$ alleviates the difficulty of $\mathcal{L}_{XI}$ by performing setwise alignment, considering the macroscopic distribution of f_i and z_i.

To balance the contributions of $\mathcal{L}_{XI}$ and $\mathcal{L}_D$, we introduce hyperparameters γ and δ to assign respective weights to each loss component during pretraining. The hybrid loss function $\mathcal{L}$ is formulated as follows:

$$\mathcal{L} = \mathcal{L}_X + \mathcal{L}_I + \gamma \mathcal{L}_{XI} + \delta \mathcal{L}_D. \tag{16}$$

After pretraining, we merge the time-series feature f_i with corresponding spectrogram feature z_i, forming a fused feature $\hat{f}_i$ for downstream tasks

$$\hat{f}_i = \alpha f_i + (1 - \alpha) z_i \tag{17}$$

where α donates the fusion factor.

III. Experiments

In this section, experiments are conducted across three scenarios: gearbox FD, valve stiction detection, and real-world industrial valve stiction detection deployment. These experiments comprehensively evaluate the effectiveness and superiority of our proposed method.

A. Datasets Description

The Gearbox dataset [22] was collected from a drivetrain dynamic simulator, as shown in Fig. 5. It comprises two distinct working conditions characterized by rotating speed–load configurations of 20-0 and 30-2. Each working condition consists of eight signals representing various parameters: 1) motor vibration (d_1); 2)–4) vibration of the planetary gearbox in x-, y-, and z-directions (d_2, d_3, and d_4); 5) motor torque (d_5); and 6)–8) vibration of the parallel gearbox in x-, y-, and z-directions (d_6, d_7, and d_8). The Gearbox dataset encompasses five distinct condition types: chipped (crack occurs in the gear teeth), miss (missing one tooth in the gear), root (crack occurs in the root of the gear teeth), surface (wear occurs on the surface of the gear), and normal (healthy).

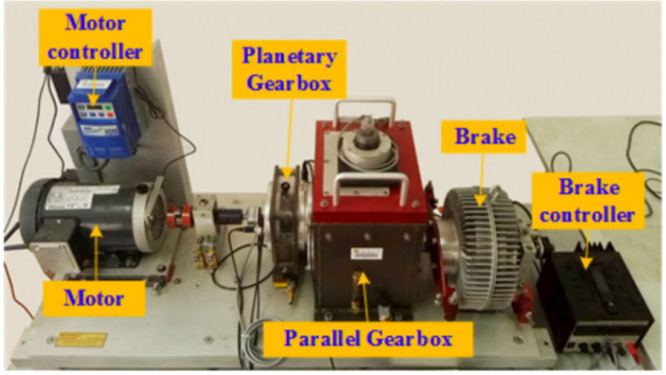


Fig. 5. Experimental setup of Gearbox dataset.

The XJTU-SY dataset [23] has emerged as a benchmark for FD research, containing complete run-to-failure data of 15 rolling element bearings. These data were obtained through systematic accelerated degradation experiments, where five bearings were subjected to three distinct speed and load rating operating conditions: 35/12, 37.5/11, and 40 Hz/10 kN. The failure of tested bearing is caused by four kinds of faults, including inner race wear (IR), cage fracture, outer race wear (OR), and outer race fracture. The tested bearing is LDK UER204, and vibration signals were collected at a sampling frequency of 26.5 kHz using two orthogonally positioned PCB 352C33 accelerometers.

The International Stiction Database (ISDB) [24] is a well-known benchmark for validating novel methods concerning control loop performance assessment. These loops were collected from various process industries, including chemical plants (CHEM), pulp and paper mills (PAP), buildings (BAS), MIN, and power plants (POW). Our goal is to determine whether the loops are stiction or not. Note that a nonstiction condition does not necessarily indicate a normal loop; it may be attributed to other issues, such as external disturbances or sensor failures.

B. Implementation Settings and Evaluation Metrics

We primarily leverage average accuracy (Acc) to evaluate the model performance, with weight-recall (w-Recall) and weight-$F1$ (w-$F1$) serving as auxiliary metrics

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{18}$$

$$\text{Prec}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{19}$$

$$w\text{-Recall} = \sum_{i=1}^{C} w_i \cdot \text{Recall}_i = \sum_{i=1}^{C} \frac{n_i}{N} \cdot \text{Recall}_i \tag{20}$$

$$w\text{-}F1 = \sum_{i=1}^{C} w_i \cdot F1_i = \sum_{i=1}^{C} \frac{n_i}{N} \cdot \frac{2 \times \text{Prec}_i \times \text{Recall}_i}{\text{Prec}_i + \text{Recall}_i} \tag{21}$$

where TP and TN represent the counts of true positive and true negative, respectively, and FP and FN indicate the counts of false positive and false negative, respectively. The subscript i indicates the class attribute c_i. C and N are the number of classes and samples, respectively. w_i and n_i signify the

TABLE I
XJTU-SY DATASET EXPERIMENTAL SETTING

Index	Condition	NC	IR	OR
1	37.5 Hz/11 kN	Bearing2_1	Bearing2_1	Bearing2_2
2	40.0 Hz/10 kN	Bearing3_4	Bearing3_3	Bearing3_5

TABLE II
TEST LOOPS DESCRIPTION AND MALFUNCTION IN THE ISDB DATASET

Loop	Type	Malfunction	Loop	Type	Malfunction
CHEM 1	*Fic*	Stiction	CHEM 24	*Fic*	Likely stiction
CHEM 2	*Fic*	Stiction	CHEM 26	*Lev*	Likely stiction
CHEM 3	*Tem*	Quantisation	CHEM 29	*Fic*	Stiction
CHEM 4	*Lev*	Tuning problem	CHEM 32	*Fic*	Likely stiction
CHEM 5	*Fic*	Stiction	CHEM 33	*Fic*	Disturbance
CHEM 6	*Fic*	Stiction	CHEM 34	*Fic*	Disturbance
CHEM 10	*Pre*	Stiction	CHEM 58	*Fic*	No oscillation
CHEM 11	*Fic*	Stiction	MIN 1	*Tem*	Stiction
CHEM 12	*Fic*	Stiction	PAP 2	*Fic*	Stiction
CHEM 13	*Ana*	Faulty sensor	PAP 4	*Con*	Deadzone
CHEM 14	*Fic*	Faulty sensor	PAP 5	*Con*	Stiction
CHEM 16	*Pre*	Interaction	PAP 7	*Fic*	Disturbance
CHEM 23	*Fic*	Likely stiction	PAP 9	*Tem*	Non-stiction

proportion factor and number of samples with ground truth c_i, respectively.

w-Recall evaluates the capacity of the model to identify all positive instances in imbalanced multicategory classification problems. w-$F1$ is a comprehensive metric representing the harmonic mean of precision and recall.

For Gearbox dataset, our experiments involved constructing two datasets, each comprising five fault types under each rotating speed–load configuration setting (20 Hz–0 V and 30 Hz–2 V). We constrained the number of samples in our train and test subsets to mimic real-world industrial scenarios, where acquiring a large quantity of labeled data may be challenging. Specifically, we randomly selected 100 samples from each class for the training subset and 200 for the test subset. Each sample has a length of 512 and a dimension of 8.

For the XJTU-SY dataset, the original dataset is not designed for FD task. Experiments were conducted by selecting normal condition (NC), IR, and OR from the 37.5-/11- and 40-Hz/10-kN working conditions. Specifically, the NC series was extracted from the normal state of the original dataset, while the IR and OR series were segmented from degradation states. The selected dataset for sampling is outlined in Table I. For each series, 500 samples were generated with a length of $L = 1024$, a stride of 1024, and a dimension of $D = 2$. Of these, 400 samples were designed as the training subset, and the remaining 100 samples constituted the testing subset.

For the ISDB dataset, we selected a total of 85 control loops with available data. For each control loop, we randomly sampled 60 samples, each comprising a length of $L = 600$ with a dimension of $D = 2$. To facilitate meaningful comparisons with other stiction detection methods, we selected 26 control loops corresponding to 1560 samples, as the test subset. Detailed information about the test subset is presented in Table II. The remaining 59 control loops (3540 samples) were allocated to the training subset. In this study, our focus is on detecting the stiction state of the loops. In line with the experimental setup in [25], a voting strategy with $N = 11$ samples is utilized to ascertain the final prediction of each loop. The pseudo-code for voting strategy is depicted in Algorithm 1.

Algorithm 1 Pseudo-Code of Voting Strategy

```
Input: The test set $\mathcal{T}_i$ of $i$-th loop with odd samples
       $N$, the trained model $AMMCL(\cdot)$
Output: The voting result $\hat{y}_i$
for $j = 1$ to $N$ do
    $y_i^j = AMMCL(\mathcal{T}_i^j)$;
end
for $k = 1$ to $C$ do
    $p_k = \sum_{j=1}^{N} \mathbb{I}(y_i^j = k)$;
end
$\hat{y}_i = \arg\max_k p_k$;
return $\hat{y}_i$
```

A two-layer transformer [26] with 16 heads, a hidden dimension of 256, and dropout ratio of 0.2 is employed as the backbone of feature extractor. The adversarial classifier and downstream classifiers are designed as three-layer MLPs with a hidden dimension of 128 and a dropout ratio of 0.2. Key model hyperparameters are set to $\alpha = 0.8$, $\gamma = 1.0$, and $\delta = 0.6$. All code was implemented in Python using PyTorch. The primary hardware includes an NVIDIA GeForce RTX 4090 GPU, an Intel Core i5-12400 CPU, and 32-GB RAM.

C. Comparison Methods

We compare the proposed AMMCL with the following state-of-the-art methods on Gearbox dataset.

1) *Deep CNN With Wide First-Layer Kernels (WD-CNN) [27]:* It uses wide kernels in the first convolutional layer to extract features and suppress high-frequency noise.
2) *CNN With Training Interference (TI-CNN) [28]:* This article presents an end-to-end method that eliminates the need for denoising preprocessing.
3) *EEMD-AlexNet [29]:* This work integrates ensemble empirical mode decomposition (EEMD) and continuous wavelet transform with AlexNet.
4) *Capsule Network (CapsNet) [30]:* It uses a novel neural network with a unique capsule unit, featuring multidimension and rich spatial information.
5) *CapsNet With an Inception Block (ICN) [31]:* This work enhances the nonlinearity of a capsule by incorporating an inception block into its architecture.
6) *Xception [32]:* This work presents an interpretation of inception modules in CNNs bridge regular and depthwise separable convolutions.
7) *Transformer [33] and Informer [34]:* These works utilize self-attention mechanisms to capture long-range

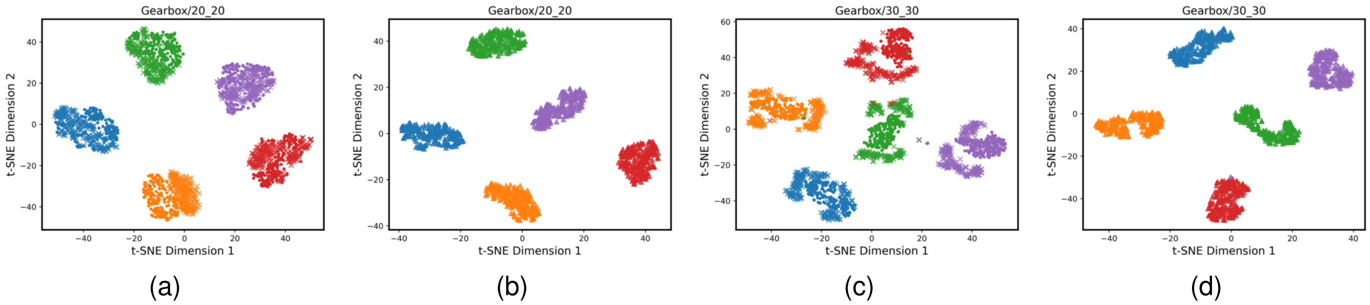


Fig. 6. t-SNE visualization of features f_i [•], z_i [×], and $\hat{f}_i$ [△] distribution of Gearbox dataset under intramodal experiments. (a) and (c) Distribution of f_i [•] and z_i [×] under 20- and 30-Hz conditions, respectively. (b) and (d) Corresponding distribution of $\hat{f}_i$ [△], respectively. (a) 20-Hz modal feature. (b) 20-Hz fusion feature. (c) 30-Hz modal feature. (d) 30-Hz fusion feature.

TABLE III

DETECTION AND COMPARISON RESULTS ON THE GEARBOX DATASET

Method	Intra-domain		Average	Cross-domain		Average	Inference Time
	20	30		20→30	30→20		
WD-CNN	92.70	91.96	92.33	82.88	83.97	83.43	4.91 ms
TI-CNN	93.83	90.54	92.19	83.61	86.73	85.17	5.76 ms
EEMD-AlexNet	93.88	90.73	92.31	81.82	81.53	81.68	7.48 ms
CapsNet	94.75	95.81	95.28	92.39	83.53	87.96	8.17 ms
ICN	96.11	94.14	95.13	88.67	91.94	90.31	8.94 ms
Transformer	93.00	98.20	95.60	95.70	82.50	89.10	8.42 ms
Informer	93.80	98.30	96.05	92.30	83.60	87.95	9.39 ms
Xception	98.70	97.92	98.31	91.34	89.78	90.56	8.72 ms
MSSLN	**99.68**	99.32	99.50	93.34	91.86	92.60	6.22 ms
AMMCL (**Ours**)	99.45	**99.72**	**99.59**	**98.19**	**96.31**	**97.25**	7.21 ms

TABLE IV

DETECTION AND COMPARISON RESULTS ON THE XJTU-SY DATASET

Method	Intra-domain		Average	Cross-domain		Average	Inference Time
	37.5	40		37.5→40	40→37.5		
WD-CNN	93.17	91.32	92.25	46.91	59.80	53.36	1.87 ms
TI-CNN	89.94	82.78	86.36	46.69	54.33	50.51	1.43 ms
EEMD-AlexNet	91.17	92.34	91.76	32.92	35.33	34.13	1.69 ms
CapsNet	88.76	87.14	87.95	42.57	48.31	45.44	2.15 ms
ICN	90.49	89.30	89.90	52.51	59.13	55.82	1.92 ms
Transformer	94.85	90.54	92.70	83.42	81.67	82.55	3.58 ms
Informer	90.21	87.07	88.64	84.05	77.99	81.02	5.04 ms
Xception	91.08	91.54	91.31	33.33	33.33	33.33	1.79 ms
MSSLN	92.10	**95.44**	93.77	33.11	60.11	46.61	1.74 ms
AMMCL (**Ours**)	**96.67**	94.22	**95.45**	**87.73**	**84.43**	**86.08**	2.96 ms

dependencies in time-series data. Informer is the improved version of transformer, which is more efficient in terms of time and space complexity.

8) *Multiscale Shared-Learning Network (MSSLN) [35]:* This work extracts and classifies fault features in vibration signals, emphasizing consistency across multiscale factors.

We compare the proposed AMMCL with the following state-of-the-art methods on ISDB dataset.

1) *Statistical-Based Method:* **Relay-based** method [36], **waveform shape** analysis [37], power spectral density, and autocorrelation function-based method (PSD/ACF) [38].
2) *Deep Learning-Based Method:* Butterfly shape-based detection method integrated with CNN (BSD-CNN) [39], multiple-timescale CNN (MTS-CNN) [14], rough stacked denoising autoencoder (RSD-AE) [40], interval probability distribution learning model (IP-DL) [41], CNN with multitimescale feature consistent constraint (MTFCC-CNN) [25], debiased contrastive learning method with expert knowledge guidance mechanism (EKGM-DCL) [2], and **Informer** [34].

D. Detection and Comparison Results

This study first presents the detection and transfer results for Gearbox and XJTU-SY datasets, along with a comparative analysis of existing FD methods. The intradomain and cross-domain experiment results, along with inference time for each sample, are presented in Table III for Gearbox dataset (with rotating speeds of 20 and 30 Hz) and in Table IV for XJTU-SY dataset (with speeds of 37.5 and 40 Hz). Additionally, Fig. 6 illustrates the feature distributions of AMMCL.

In intradomain experiments, the training and test subsets are derived from data under the same rotating speed condition. For cross-domain experiments, the notation 20 $\rightarrow$ 30 Hz indicates that the model is trained on a 20-Hz training subset and fine-tuned with part of the 30-Hz training subset, with performance tested on the 30-Hz test subset. In intradomain experiments, AMMCL achieves an Acc of 99.45% at 20 Hz and 99.72% at 30 Hz. Fig. 6(a) and (c) illustrates the distributions of f_i [•] and z_i [×]. Different categories of f_i [•] and z_i [×] form distinct clustered, while features within the same class show a clustered distribution. The fusion features $\hat{f}_i$ [△] [see Fig. 6(b) and (d)] also display sharply defined boundaries. Compared to MSSLN, AMMCL achieves superior performance at 30 Hz and better overall average performance on the Gearbox dataset. AMMCL outperforms most other methods in Acc performance and inference time, with only MSSLN showing competitive results.

Cross-domain experiments assessed the generalization of the extracted features. Despite a slight drop in Acc compared to intramodal experiments, AMMCL still outperformed other methods in feature generalization, achieving Acc of 98.19% and 96.31% under 30- and 20-Hz conditions. This demonstrates the AMMCL's strong feature extraction and potential for industrial deployment. As shown in Fig. 7(a) and (c), f_i [•] has a tighter distribution, while z_i [×] is more scattered.

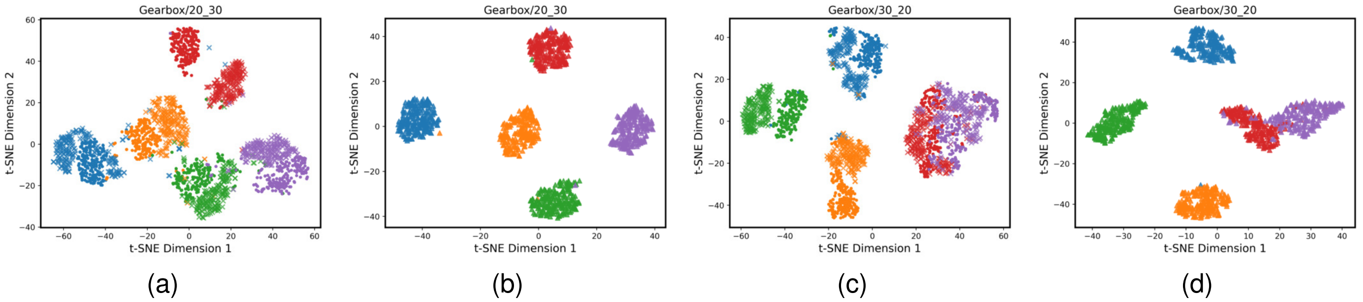


Fig. 7. t-SNE visualization of features f_i [•], z_i [×], and $\hat{f}_i$ [△] distribution of Gearbox dataset under cross-modal experiments. (a) and (c) Distribution of f_i [•] and z_i [×] under 20- → 30- and 30- → 20-Hz conditions, respectively. (b) and (d) Corresponding distribution of $\hat{f}_i$ [△], respectively. (a) 20- → 30-Hz modal feature. (b) 20- → 30-Hz fusion feature. (c) 30- → 20-Hz modal feature. (d) 30- → 20-Hz fusion feature.

However, $\hat{f}_i$ [△] [see Fig. 7(b) and (d)] remains well-defined, highlighting great cross-domain performance.

Overall, AMMCL achieves exceptional results. Specifically, compared to the transformer-based methods (i.e., transformer and Informer), it can achieve more intrinsic features for cross-domain FD tasks with lower inference time, attributed to its cross-modal contrastive learning framework and SSP operation. Unlike MSSLN, Xception, ICN, and so on, which build the complex unimodal feature extractors, AMMCL aims to learn intrinsic, modal-invariant features from the time series and spectrogram, making it more effective in cross-domain scenarios. Additionally, AMMCL's inference time of 7.21 ms strikes a good balance between efficiency and Acc.

In intradomain experiments on the XJTU-SY dataset, AMMCL achieves exceptional Acc, reaching 96.67% at 37.5 Hz and 94.22% at 40 Hz, with an average of 95.45%, which surpasses all comparative methods, including MSSLN (93.77%) and transformer (92.70%). For cross-domain tasks, AMMCL maintains robust generalization: under 37.5- → 40- and 40- → 37.5-Hz configurations, it attains 87.73% and 84.43% Acc, respectively, yielding an average cross-domain Acc of 86.08%. This significantly outperforms transformer (82.55%) by 3.53% and MSSLN (46.61%) by 39.47%, highlighting its superior capability to extract modal-invariant features across rotational speed shifts. While AMMCL's inference time (2.96 ms) is slightly higher than lightweight CNNs (e.g., TI-CNN: 1.43 ms), it remains 17.3% faster than transformer (3.58 ms) and 41.3% faster than informer (5.04 ms), striking an optimal balance between Acc and efficiency. XJUT-SY dataset contains substantial industrial noise. Cross-domain behavior needs to resist the influences of bearing load and rotational speed differences, which can effectively reflect the robustness of model features. Overall, AMMCL demonstrates the state-of-the-art performance on XJTU-SY, validating its strong industrial applicability for FD under variable operational conditions.

For the ISDB dataset, model performance is evaluated via voting Acc for loop stiction detection. Each control loop's state is determined via aggregating classification results from its samples. Comparative results in Table V show that AMMCL achieves voting Acc of 0.8846 (23/26), outperforming ten other algorithms. Furthermore, it surpasses traditional statistic methods, such as relay-based, waveform shape, and PSD/ACF methods, as well as deep learning methods, such as IP-DL, MTFCC-CNN, EKGM-DCL, and informer, correctly detecting 23 loops. MTFCC-CNN converts time series into images and use multikernel CNN to extract semantic features. EKGM-DCL introduces expert knowledge to guide the model training. However, their contributions still focus on enhancing models within unimodal scenarios, without breaking the shackles of unimodal-based augmentation techniques and only extract the shallow invariant information.

TABLE V
DETECTION AND COMPARISON RESULTS ON THE ISDB DATASET

Method	Voting Accuracy	Voting Results
Relay-based method	0.6538	17/26
Waveform shape analysis	0.4231	11/26
PSD/ACF method	0.6923	18/26
BSD-CNN method	0.7692	20/26
MTS-CNN method	0.8076	21/26
RSD-AE method	0.6538	17/26
IP-DL method	0.8076	21/26
Informer method	0.8076	21/26
MTFCC-CNN method	0.8461	22/26
EKGM-DCL method	0.8461	22/26
AMMCL (**Ours**)	**0.8846**	**23/26**

E. Actual Industrial Valve Detection Deployment

In practical applications, the proposed AMMCL is verified using control loops from both a hardware experimental system and actual industrial environments.

The hardware experimental system comprises a liquid-level control loop (LIC201) and two flow control loops (FIC201 and FIC202). As shown in Fig. 8, the system includes valves (V201, V202, and V203), electromagnetic flow meters (M201, M202, and M203), and a pressure sensor (L201). L201 measures the bottom pressure of Tank 202 and converts it to a liquid level reading. V203 and V202 control the water inflow into Tank 202, while V201 controls the outflow.

The other three control loops (i.e., PIC23002, FIC3107, and F6304) are collected from real industrial environments. PIC23002 is a pressure control loop affected by unknown

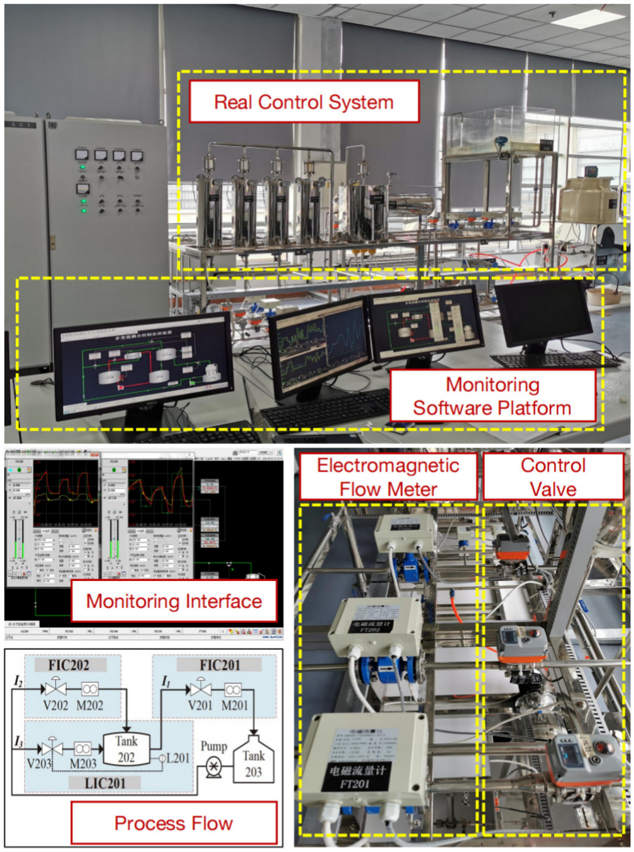


Fig. 8. Control loop condition monitoring experimental system.

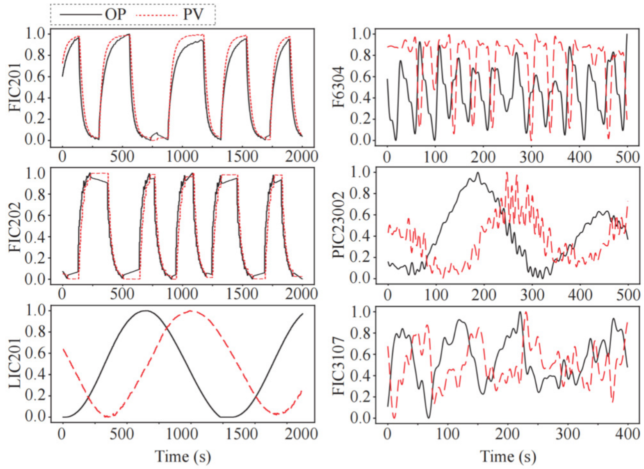


Fig. 9. Raw data of six real control loops.

external disturbances. FIC3107 is a normal flow control loop. F6304 is a flow control loop with recorded stiction. Fig. 9 shows the raw data of these six loops. The X-axis indicates time in seconds, and the Y-axis shows normalized values ranging from 0 to 1. For flow control loops FIC201, FIC202, FIC3017, and F6304, the unit is m^3/h. LIC201 uses cm, and PIC23002 uses kPa.

TABLE VI
DEPLOYMENT RESULTS ON ACTUAL HARDWARE SYSTEMS

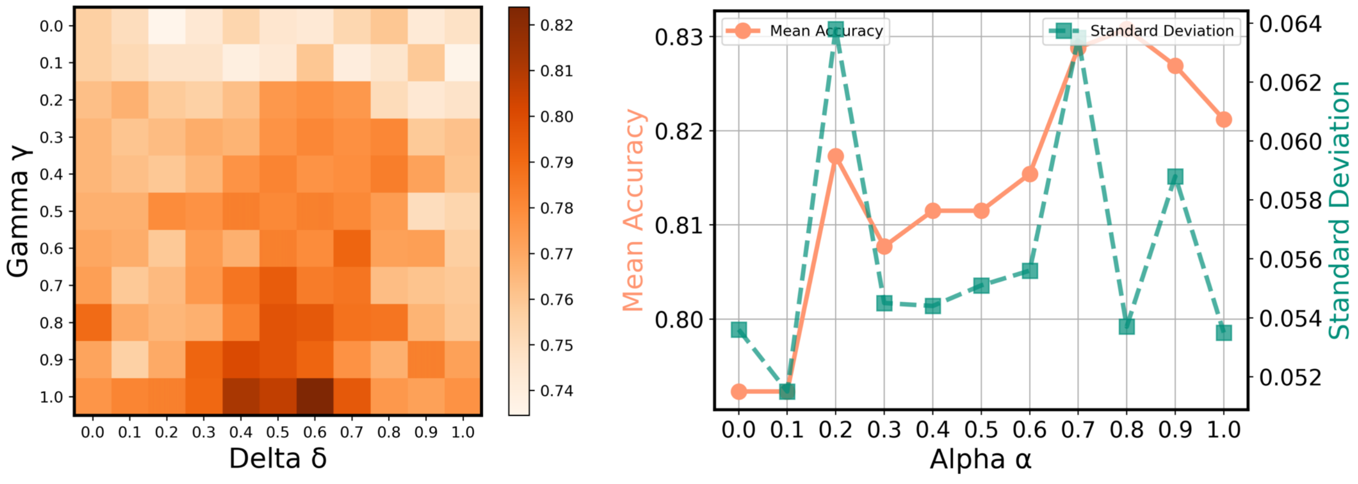
Method	PIC23002	FIC3107	F6304	FIC201	FIC202	LIC201
Ground Truth	External Distribution	Normal	Stiction	Normal	Normal	Normal
SVM	Nonstic	Nonstic	Nonstic	Stic	Nonstic	Nonstic
XgBoost	Nonstic	Nonstic	Stic	Stic	Nonstic	Nonstic
LeNet-5	Stic	Stic	Stic	Nonstic	Stic	Nonstic
MTFCC-CNN	Nonstic	Nonstic	Stic	Nonstic	Nonstic	Stic
AMMCL	Nonstic	Nonstic	Stic	Nonstic	Nonstic	Nonstic

To evaluate AMMCL's applicability in real industrial settings, the ISDB-trained model was directly applied to detect actual industrial valve faults without fine-tuning. Deployment results are in Table VI, with underlined entries indicating false detections. AMMCL accurately detected stiction in all loops, outperforming statistic-based (SVM and XgBoost) and deep learning-based (LeNet-5 and MTFCC-CNN) methods. Despite the solid theoretical foundation and wide industrial use of SVM and XgBoost, and the strong performance of MTFCC-CNN and LeNet-5 in valve stiction detection, AMMCL proved most effective. The comparison confirms AMMCL's feasibility for real industrial cross-domain deployment.

F. Key Hyperparameter Selection

The proposed AMMCL framework's performance is significantly influenced by several crucial hyperparameters, including the weighted coefficients γ and δ of the hybrid loss function $\mathcal{L}$ in (16) and the fusion factor α in (17). To illustrate their impact and identify optimal values, we conducted experiments on the ISDB dataset. To streamline the experiments, we assessed the relative significance of these hyperparameters and sequentially froze them in the following order.

1) γ and δ Orthogonal Experiments: Following (16), the weighting factors γ and δ balance the contributions of $\mathcal{L}_{XI}$ and $\mathcal{L}_D$ within the hybrid loss $\mathcal{L}$. Here, $\mathcal{L}_{XI}$ performs cross-modal contrastive learning to capture intrinsic and generalizable status message of the raw data, while $\mathcal{L}_D$ reduces modal-specific information remained in the cross-modal features f_i and z_i due to inherent modal differences. The synergy between these two loss functions is crucial for improving the model's generalization ability from both elementwise and setwise perspectives. The orthogonal experiments are depicted in Fig. 10(a). The heatmap illustrates a gradient of colors from the left and right sides toward the middle and from top to bottom. The results presented in Fig. 10(a) reveal that a larger value of γ (weight of $\mathcal{L}_{XI}$) and a medium value of δ (weight of $\mathcal{L}_D$) ensure that AMMCL achieves the highest Acc. This is because a larger γ emphasizes the role of $\mathcal{L}_{XI}$, enabling the model to effectively align features at an element-to-element fine semantic matching with stronger constraints. On the other hand, a medium δ allows $\mathcal{L}_D$ to provide a moderate level of macroscopic constraints, further bridging the semantic spaces of the two modalities and enhancing the modal-invariant information. This balance between detailed semantic matching and macroscopic semantic space alignment leads to better optimization results, enabling the model to effectively combine the strengths of both modalities for more



(a) (b)

Fig. 10. Key hyperparameter selection experiments. (a) Heat map of the orthogonal experiments for γ and δ, where AMMCL Acc increases with darker colors. (b) AMMCL Acc and standard deviation across different α values. The left vertical axis corresponds to mean Acc (orange bars), while the right vertical axis corresponds to standard deviation (green bars).

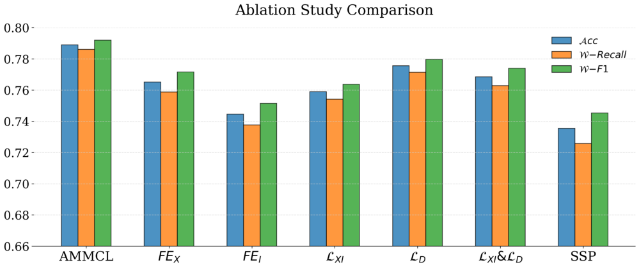


Fig. 11. Ablation study experiments on ISDB dataset. Except for the AMMCL column, the names in the other columns indicate the modules that were blocked during the experiments.

accurate and reliable FD. Finally, we select $\gamma = 1.0$ and $\delta = 0.6$.

2) α Experiments: The parameter α governs the interplay between time-series feature f_i and spectrogram feature z_i in the fused feature $\hat{f}_i$, both of which are extracted via the adversarial training strategy. To determine the optimal value of α, we conducted an ablation study varying α from 0.0 to 1.0, as shown in Fig. 10. Notably, both f_i and z_i independently achieve decent classification performance. However, when $\alpha = 0.8$, the synergy between is optimized, enabling AMMCL to reach the highest average Acc of 0.835. We therefore set $\alpha = 0.8$ to maximize the combined their combined potential.

G. Ablation Experiments

The AMMCL framework comprises five key modules: time-series encoder (FE$_X$), spectrogram encoder (FE$_I$), cross-modal InfoNCE loss ($\mathcal{L}_{XI}$), Adversarial CrossEntropy Loss ($\mathcal{L}_D$), and SSP. In this section, we use ablation experiments on the ISDB dataset to demonstrate each module's effectiveness. We randomly select 20 loops as test subset and use the remaining loops for training, without applying voting operations for statistical results. Fig. 11 shows experiments where different modules are masked based on the full AMMCL framework, highlighting each module's specific contribution. w-$F1$, a metric that balances both false alarms and omissions, is crucial for unbalanced classification tasks and serves as our primary evaluation metric.

For the FE$_X$ and FE$_I$ columns, we evaluate the detection performance of utilizing time series and spectrogram individually to validate the necessity of feature fusion. Results demonstrate that spectrogram alone achieves w-$F1 = 0.7717$, outperforming time series alone with w-$F1 = 0.7516$. This indicates spectrogram better reflects system states, as system state changes are more discernible in spectrogram patterns.

For the columns $\mathcal{L}_{XI}$, $\mathcal{L}_D$, and $\mathcal{L}_{XI}$ & $\mathcal{L}_D$, we investigate the contribution of cross-modal InfoNCE loss $\mathcal{L}_{XI}$ and adversarial cross-entropy loss $\mathcal{L}_D$ to feature fusion. When $\mathcal{L}_{XI}$ is masked ($\mathcal{L}_{XI}$ column), AMMCL achieves a w-$F1$ of 0.7638, surpassing the results of $\mathcal{L}_D$ and $\mathcal{L}_{XI}$ & $\mathcal{L}_D$ columns. If only $\mathcal{L}_D$ is active in AMMCL, f_i and z_i focus on extracting modal-specific information, leading to a lack of relevance and weakened feature fusion effectiveness and rationality. For the $\mathcal{L}_D$ column, $\mathcal{L}_D$ is masked, resulting in f_i and z_i primarily containing modal-invariant information. This achieves better results than $\mathcal{L}_{XI}$ column (w-$F1 = 0.7638$) but still lags behind AMMCL column (w-$F1 = 0.7921$) due to the lack of assistance from setwise margin reduction. In the $\mathcal{L}_{XI}$ & $\mathcal{L}_D$ column, f_i and z_i are fused without constraints, yielding results (w-$F1 = 0.7741$) that lie between $\mathcal{L}_{XI}$ and $\mathcal{L}_D$ columns. The synergy of $\mathcal{L}_{XI}$ and $\mathcal{L}_D$ is proven effective.

For the SSP column, the contribution of SSP for the extraction of f_i is researched. The slicewise (AMMCL column) and pointwise (SSP column) methods represent different approaches to handling time series. Pointwise methods focus on individual timestamps and their relationship, which limits their ability to decompose and recognize modes in time series. In contrast, slicewise methods, which treat time slices a basic processing units, can better capture varying patterns with statistical meaning. The pointwise AMMCL (SSP column) yields the worst results (w-$F1 = 0.7454$), even lower than the FE_X and FE_I alone. Meanwhile, slicewise AMMCL achieves a w-$F1$ improvement of 0.0467 over the SSP column.

Ablation studies confirm the effectiveness of each module in the AMMCL framework and demonstrate their synergistic effect in enhancing the overall performance.

IV. Conclusion

This article proposes a novel AMMCL framework for FD. It tackles the dilemma of intramodal SSCL methods prone to learning shallow, domain-dependent features. SSP based on dominant frequency treats each time slice as the model's fundamental processing unit. This expands the receptive field of the model and endowing its ability to recognize diverse time-series patterns with statistical meanings. Multimodal contrastive learning between time series and spectrograms is proposed. It aligns elementwise positive pairs across modalities, enabling the model to capture more robust, intrinsic, and modal-invariant features. This method overcomes the limitations of traditional intramodal contrastive learning algorithms based on augmentation techniques, which extract surface invariant part across views. Furthermore, an adversarial training strategy is incorporated to reduce setwise margin between two modal features, aiding multimodal contrastive learning. Experiments on the Gearbox and XJTU-SY datasets validate AMMCL's effectiveness in both intradomain and cross-domain settings. ISDB dataset experiments show its superiority over ten other stiction detection algorithms. Moreover, the deployment of the trained AMMCL on actual industrial valve detection has achieved the best performance. Ablation studies confirm each module's effectiveness within AMMCL. Future work will explore why the multimodal contrastive learning clusters the same-class features, while intramodal contrastive learning distributes them uniformly.

References

[1] P. N. Mueller, L. Woelfl, and S. Can, "Bridging the gap between AI and the industry—A study on bearing fault detection in PMSM-driven systems using CNN and inverter measurement," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106834.

[2] K. Zhang, R. Cai, C. Zhou, and Y. Liu, "Debiased contrastive learning for time-series representation learning and fault detection," *IEEE Trans. Ind. Informat.*, vol. 20, no. 5, pp. 7641–7653, May 2024.

[3] J. B. Thomas and K. V. Shihabudheen, "Neural architecture search algorithm to optimize deep transformer model for fault detection in electrical power distribution systems," *Eng. Appl. Artif. Intell.*, vol. 120, Apr. 2023, Art. no. 105890.

[4] R. Zemouri, R. Ibrahim, and A. Tahan, "Hydrogenerator early fault detection: Sparse dictionary learning jointly with the variational autoencoder," *Eng. Appl. Artif. Intell.*, vol. 120, Apr. 2023, Art. no. 105859.

[5] R. Cai, L. Peng, Z. Lu, K. Zhang, and Y. Liu, "DCS: Debiased contrastive learning with weak supervision for time series classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 5625–5629.

[6] R. Cai, W. Gao, L. Peng, Z. Lu, K. Zhang, and Y. Liu, "Debiased contrastive learning with supervision guidance for industrial fault detection," *IEEE Trans. Ind. Informat.*, vol. 20, no. 11, pp. 12814–12825, Nov. 2024.

[7] Y. Zhang, Z. Liu, and Q. Huang, "A contrastive learning-based fault diagnosis method for rotating machinery with limited and imbalanced labels," *IEEE Sensors J.*, vol. 23, no. 14, pp. 16402–16412, Jul. 2023.

[8] C. Azuma, T. Ito, and T. Shimobaba, "Adversarial domain adaptation using contrastive learning," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106394.

[9] Y. Liu, W. Wen, Y. Bai, and Q. Meng, "Self-supervised feature extraction via time–frequency contrast for intelligent fault diagnosis of rotating machinery," *Measurement*, vol. 210, Jan. 2023, Art. no. 112551.

[10] K. Zhang et al., "Self-supervised learning for time series analysis: Taxonomy, progress, and prospects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6775–6794, Oct. 2024.

[11] P. Peng et al., "Progressively balanced supervised contrastive representation learning for long-tailed fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[12] R. Chen, Z. Cai, and J. Yuan, "UIESC: An underwater image enhancement framework via self-attention and contrastive learning," *IEEE Trans. Ind. Informat.*, vol. 19, no. 12, pp. 11701–11711, Dec. 2023.

[13] P. Peng, J. Lu, T. Xie, S. Tao, H. Wang, and H. Zhang, "Open-set fault diagnosis via supervised contrastive learning with negative out-of-distribution data augmentation," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 2463–2473, Mar. 2023.

[14] K. Zhang, Y. Liu, Y. Gu, X. Ruan, and J. Wang, "Multiple-timescale feature learning strategy for valve stiction detection based on convolutional neural network," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 3, pp. 1478–1488, Jun. 2022.

[15] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5549–5560, May 2023.

[16] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," 2021, *arXiv:2110.02642*.

[17] Y. Chen, K. Ren, Y. Wang, Y. Fang, W. Sun, and D. Li, "Contiformer: Continuous-time transformer for irregular time series modeling," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023, pp. 1–11.

[18] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–24.

[19] X. Yuan et al., "Multimodal contrastive training for visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6995–7004.

[20] S. Mai, Y. Zeng, and H. Hu, "Learning from the global view: Supervised contrastive learning of multimodal representation," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101920.

[21] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, "A comprehensive survey on contrastive learning," *Neurocomputing*, vol. 610, Dec. 2024, Art. no. 128645.

[22] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.

[23] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2020.

[24] M. Jelali and B. Huang, *Detection and Diagnosis of Stiction in Control Loops: State of the Art and Advanced Methods*. London, U.K.: Springer, 2009.

[25] K. Zhang, Y. Liu, Y. Gu, J. Wang, and X. Ruan, "Valve stiction detection using multitimescale feature consistent constraint for time-series data," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 3, pp. 1488–1499, Jun. 2023.

[26] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-variation modeling for general time series analysis," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–23.

[27] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, Feb. 2017.

[28] 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.

[29] X. Shi, Y. Cheng, B. Zhang, and H. Zhang, "Intelligent fault diagnosis of bearings based on feature model and alexnet neural network," in *Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM)*, Jun. 2020, pp. 1–6.

[30] L. Chen, N. Qin, X. Dai, and D. Huang, "Fault diagnosis of high-speed train bogie based on capsule network," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6203–6211, Sep. 2020.

[31] Z. Zhu, G. Peng, Y. Chen, and H. Gao, “A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis,” *Neurocomputing*, vol. 323, pp. 62–75, Jan. 2019.

[32] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[33] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[34] H. Zhou et al., “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 12, pp. 11106–11115.

[35] Z. Chen, S. Tian, X. Shi, and H. Lu, “Multiscale shared learning for fault diagnosis of rotating machinery in transportation infrastructures,” *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 447–458, Jan. 2023.

[36] M. Rossi and C. Scali, “A comparison of techniques for automatic detection of stiction: Simulation and application to industrial data,” *J. Process Control*, vol. 15, no. 5, pp. 505–514, Aug. 2005.

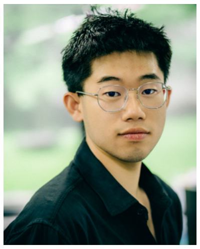
[37] A. Singhal and T. I. Salsbury, “A simple method for detecting valve stiction in oscillating control loops,” *J. Process Control*, vol. 15, no. 4, pp. 371–382, Jun. 2005.

[38] S. Karra and M. N. Karim, “Comprehensive methodology for detection and diagnosis of oscillatory control loops,” *Control Eng. Pract.*, vol. 17, no. 8, pp. 939–956, Aug. 2009.

[39] B. Kamaruddin, H. Zabiri, A. A. A. M. Amiruddin, W. K. Teh, M. Ramasamy, and S. S. Jeremiah, “A simple model-free butterfly shape-based detection (BSD) method integrated with deep learning CNN for valve stiction detection and quantification,” *J. Process Control*, vol. 87, pp. 1–16, Mar. 2020.

[40] M. Khodayar, O. Kaynak, and M. E. Khodayar, “Rough deep neural architecture for short-term wind speed forecasting,” *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 2770–2779, Dec. 2017.

[41] M. Khodayar, J. Wang, and M. Manthouri, “Interval deep generative neural network for wind speed forecasting,” *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3974–3989, Jul. 2019.



Rongyao Cai received the B.Eng. degree in chemical engineering from Zhejiang University of Technology, Hangzhou, China, in 2022. He is currently pursuing the Ph.D. degree in control engineering with the College of Control Science and Engineering, Zhejiang University, Hangzhou.

His major research interests include time-series analysis, deep learning, and domain adaptation.

Kexin Zhang (Member, IEEE) received the B.Eng. and M.Eng. degrees in control engineering from China University of Geosciences, Wuhan, China, in 2016 and 2019, respectively, and the Ph.D. degree in control engineering and science from Zhejiang University, Hangzhou, China, in 2023.

He is currently an Assistant Professor with Huzhou Institute of Zhejiang University, Huzhou, China. He has authored more than ten articles, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and IEEE/ASME TRANSACTIONS ON MECHATRONICS. His major research interests include intelligent time-series analysis, deep learning, data-driven industrial fault diagnosis, and artificial intelligence security.

Hanchen Tai received the B.Eng. degree in ocean engineering from Zhejiang University, Hangzhou, China, in 2022, where he is currently pursuing the M.Eng. degree in control engineering with the College of Control Science and Engineering.

His major research interests include deep learning in 3-D object detection and multimodal 3-D scene understanding.

Yang Zhou received the B.S. degree in automation and the Ph.D. degree in control science and engineering from China University of Geosciences, Wuhan, China, in 2016 and 2022, respectively.

He was the Joint Ph.D. Student with Tokyo University of Technology, Tokyo, Japan, from 2019 to 2020. He joined China University of Geosciences in 2023, where he currently holds a post-doctoral position with the School of Automation. His research interests include data-driven modeling and optimization algorithm.

Yuanyuan Ding received the B.Eng. degree in control engineering from Shandong University, Jinan, China, in 2022. She is currently pursuing the M.Eng. degree in control engineering with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China.

Her major research interests include machine learning and data mining, computer vision, and 3-D reconstruction.

Chunlin Zhou received the B.S. and M.Eng. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2006, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2012, all in mechatronics engineering.

He is currently an Associate Professor with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. He is also the Deputy Sectary of the University Students Robotics Competition Committee of Zhejiang Province. He has authored more than 30 academic articles and the inventor of five Chinese/U.S. patents. His research interests include mechatronics design, motion control, and their applications in surgical robots and biomimetic robots.

Yong Liu (Member, IEEE) received the B.S. degree in computer science and engineering and the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2001 and 2007, respectively.

He is currently a Professor with the Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University. He has authored or co-authored more than 100 research papers in machine learning, computer vision, information fusion, and robotics, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, *ACM Transactions on Graphics*, IEEE TRANSACTIONS ON ROBOTICS, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, ICLR, NeurIPS, AAAI, CVPR, ECCV, ICCV, IROS, and ICRA. His current research interests include machine learning, robotics vision, information processing, and granular computing.