# Shared-Exclusive Disentangled Representation Learning With Channel Attention for Incipient Fault Detection in Multimode Power Inverter Systems

Shuihui Rao, *Graduate Student Member, IEEE*, Yu Shuai, *Graduate Student Member, IEEE*, Zhihua Xiong, *Member, IEEE*, and Min Wang, *Member, IEEE*

***Abstract*—Inverter systems play a critical role in modern power conversion and energy management applications, where they often operate under diverse and complex modes. The presence of multiple operating modes leads to varying data distributions, posing significant challenges for accurate fault detection across different working scenarios. Existing methods struggle to disentangle mode-invariant and mode-specific features that effectively represent incipient faults, while also lacking selective attention across the disentangled feature spaces. To address this issue, this article proposes a shared and exclusive representation learning model (SERLM) for fault detection in multimode inverter systems. SERLM leverages shared and exclusive modules to extract mode-invariant and mode-specific features, respectively, and employs a novel multiscale regularized channel attention strategy to enhance feature fusion. In addition, two statistics are designed: one based on shared features to detect global anomalies across modes and another derived from reconstruction error to capture localized deviations. The effectiveness of feature extraction in SERLM is rigorously analyzed from a theoretical perspective. The proposed approach is validated through experiments on a power inverter system, demonstrating its superiority in detecting incipient faults under multiple modes, achieving a fault detection rate that is 6.25% higher than state-of-the-art methods.**

***Index Terms*—Fault detection, inverter systems, multimode processes, representation learning.**

## I. Introduction

INVERTER systems have evolved into critical infrastructure components in modern power grids, enabling efficient dc to ac (dc/ac) conversion [1], [2], [3], [4]. Notably, driven by varying load demands, environmental factors, and integration with renewable energy sources, these systems are often required to operate across multiple modes [5], [6]. While the multimode characteristics enhance system adaptability, they introduce significant challenges in incipient fault detection due to overlapping characteristics between faults and normal operational variations [7].

In practical scenarios, inverter systems often operate under multiple modes [6]. A variety of fault detection approaches for multimode processes have been proposed, typically categorized into single-model and multimodel methods. The former establishes a single model to monitor all operating conditions. Li et al. [8] adapted the model to mode variations through recursive updates. Zhang et al. [9] projected multimode data onto a common subspace and several mode-specific subspaces. Wang et al. [10] developed a ProbSparse attention-based transformer (PSAT) that improves the key general feature extraction and enables real-time fault diagnosis under varying working conditions. Multimodel methods establish a separate model for each mode. Sharifi et al. [11] proposed a sensor fault diagnosis methodology for nonlinear systems using mixture of probabilistic principal component analysis (MPPCA) and Bayesian analysis of residuals. Xu et al. [12] proposed a decision self-regulating network (DSRN) that employs a self-regulation mechanism to enhance fault diagnosis across different operating conditions.

Inverter system failures often stem from incipient faults, which are particularly challenging to detect [4]. In single-mode processes, incipient fault characteristics are often subtle and easily obscured by noise, which reduces the effectiveness of most data-driven methods. Deep learning-based approaches address this challenge by leveraging neural networks to automatically extract high-dimensional features. Cheng et al. [13] developed a manifold learning approach using a local linear generalized autoencoder for incipient fault detection. Wang et al. [14] proposed an autoencoder-based multiscale temporal convolutional network that extracts multiscale temporal features and long-term dependencies from time series. In contrast, statistical analysis-based approaches typically construct monitoring statistics from process data. Shang et al. [15] monitored relative changes in

Received 27 June 2025; revised 11 October 2025; accepted 18 November 2025. Date of publication 24 November 2025; date of current version 25 February 2026. This work was supported in part by the National Key R&D Program of China under Grant 2021YFC2902701, in part by the National Natural Science Foundation of China under Grant 62303090, in part by the Postdoctoral Science Foundation of China under Grant 2023M740516, and in part by the Natural Science Foundation of Sichuan Province under Grant 2024NSFSC1480. Recommended for publication by Associate Editor M. Wang. *(Corresponding author: Min Wang.)*

Shuihui Rao and Zhihua Xiong are with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: rsh23@mails.tsinghua.edu.cn; zhxiong@tsinghua.edu.cn).

Yu Shuai and Min Wang are with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: 202322060338@std.uestc.edu.cn; mwang@uestc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPEL.2025.3636495.

Digital Object Identifier 10.1109/TPEL.2025.3636495

ordered eigenvalues of covariance matrices to detect incipient faults. Liu et al. [16] proposed a feature ensemble network (FENet), which integrates outputs from basic detectors and applies sliding-window patching and PCA to construct a discriminative feature representation.

However, these works [13], [14], [15], [16] face challenges when applied to multimode processes. Although the conventional methods for multimode processes have achieved significant progress [8], [9], [10], [11], [12], they struggle to disentangle mode-invariant and mode-specific features that effectively represent incipient faults and lack selective attention across the disentangled feature spaces. Therefore, this article proposes a shared and exclusive representation learning model (SERLM) for incipient fault detection in multimode inverter systems. The proposed method employs cycle-consistent generative adversarial networks (CycleGAN) [17] to transform between the raw data domain and the corresponding FENet-extracted feature domain, facilitating the extraction of latent features that represent incipient faults. The generators and discriminator in SERLM consist of shared and exclusive modules, which extract shared features across all modes and exclusive features for each mode, respectively. Moreover, a novel channel attention strategy is utilized to integrate shared and exclusive features. The main contributions are outlined as follows.

1) *A novel method (SERLM) is proposed to disentangle mode-invariant and mode-specific features and applies selective channel attention to enhance the representation of incipient faults.* SERLM utilizes shared and exclusive modules to extract corresponding features, which are subsequently fused to obtain intrinsic features representing incipient faults. In addition, shared feature discriminators are introduced to ensure the generalizability of shared features across all modes, while a feature orthogonality loss is incorporated to enforce the independence between shared and exclusive features.
2) *A multiscale regularized channel attention (MRCA) strategy is proposed to improve feature fusion.* By leveraging multiscale convolution and smooth attention regularization, MRCA adaptively emphasizes informative feature channels while suppressing irrelevant ones, thereby improving the robustness of incipient fault representation.
3) *The effectiveness of feature extraction in SERLM is rigorously analyzed from a theoretical perspective.* The feature regularization loss in SERLM is shown to drive redundant components to diminish with high probability. Furthermore, the identifiability of shared features is established, demonstrating that they are identifiable up to an orthogonal transformation.

The rest of this article is organized as follows. Section II introduces the theoretical background. Section III describes the proposed method in detail. The experimental evaluation is presented in Section IV. Finally, Section V concludes this article.

## II. Related Theory

### A. Cycle-Consistent Generative Adversarial Networks (GANs)

CycleGAN [17] is a GAN proposed for unsupervised domain translation. It introduces a cycle consistency loss that ensures the reconstructed data are consistent with the original input, thereby enabling effective bidirectional domain mapping

$$\mathcal{L}_{\text{cycle}}(G_1, G_2) = \mathbb{E}_{x\sim p_X}\left[\|G_2(G_1(x)) - x\|_1\right] \tag{1}$$

where $G_1$ and $G_2$ are the two generators. Here, CycleGAN is employed for the transformation between the raw data domain and the feature domain generated by FENet.

### B. Channel Attention

Channel attention is a mechanism that selectively emphasizes informative channels while suppressing less relevant ones in convolutional neural networks. A representative approach is efficient channel attention [18], which computes a channelwise attention map by aggregating feature information across spatial dimensions

$$\mathbf{a} = \sigma\left(\text{Conv1D}_s(\mathbf{z})\right) \tag{2}$$

where $\mathbf{z}$ is the global average pooled channel descriptor, $\text{Conv1D}_s(\cdot)$ denotes a 1-D convolution with kernel size $s$, and $\sigma(\cdot)$ is the sigmoid activation. This attention map is subsequently utilized to reweight the feature maps, enabling the model to focus on the most relevant features for the task.

## III. Methodology

In this section, the proposed SERLM is comprehensively introduced for fault detection in multimode processes. SERLM aims to learn shared and exclusive features across different modes and to adaptively integrate them through attention-based fusion to construct discriminative representations for incipient faults. The overall architecture of SERLM, as shown in Fig. 1, consists of a mode classifier ($C_M$), two generators ($G_X$ and $G_Y$), a domain discriminator ($D_Y$), and three shared feature discriminators ($D^s_{G_X}$, $D^s_{G_Y}$, and $D^s_{D_Y}$). The generators and domain discriminator consist of a shared feature extractor, an exclusive feature extractor, an MRCA module, and a feature fusion module. The detailed architecture of each component is illustrated in Fig. 2.

SERLM employs CycleGAN to model the transformation between raw input data and the corresponding features extracted by FENet. Within its generators and domain discriminator, a novel shared-exclusive disentangled learning framework is introduced to separate mode-invariant and mode-specific features. Compared with the original CycleGAN, SERLM incorporates shared feature discriminators to ensure the consistency of the extracted mode-invariant features.

### A. Multiscale Regularized Channel Attention

The MRCA is designed to effectively integrate shared and exclusive features, as illustrated in Fig. 3. MRCA enhances informative channels while suppressing redundant ones by means of multiscale convolutional attention and a regularization term.

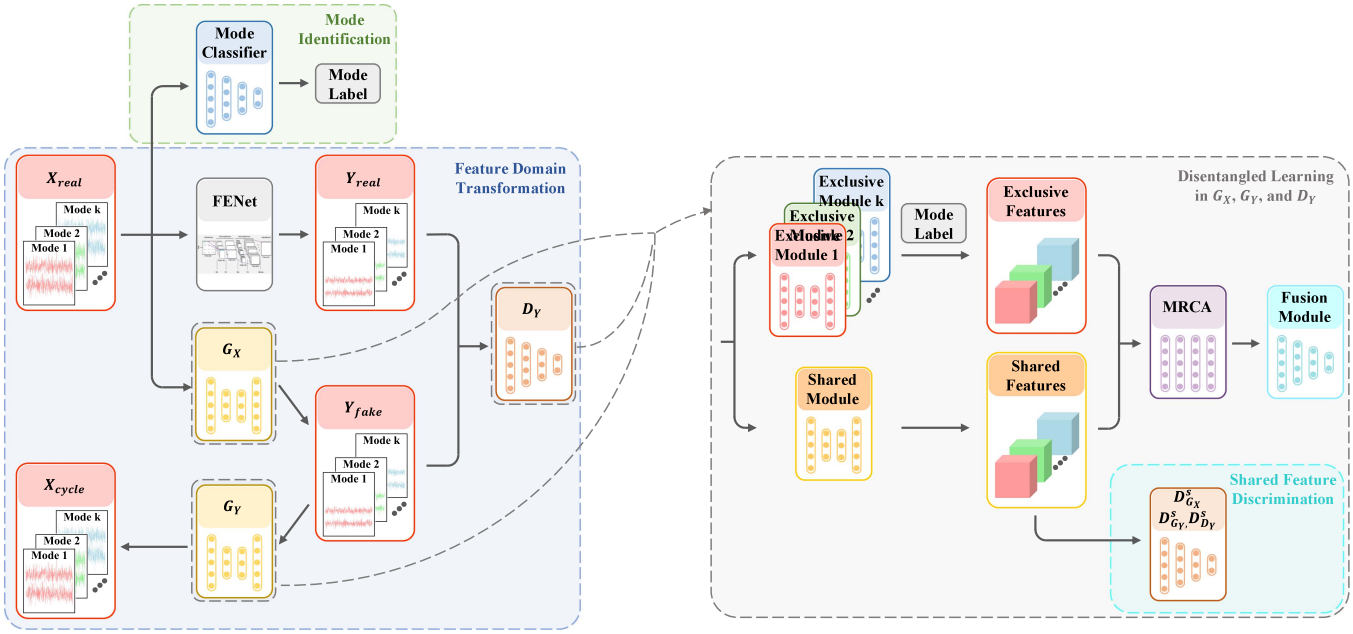



Fig. 1. Architecture of SERLM.

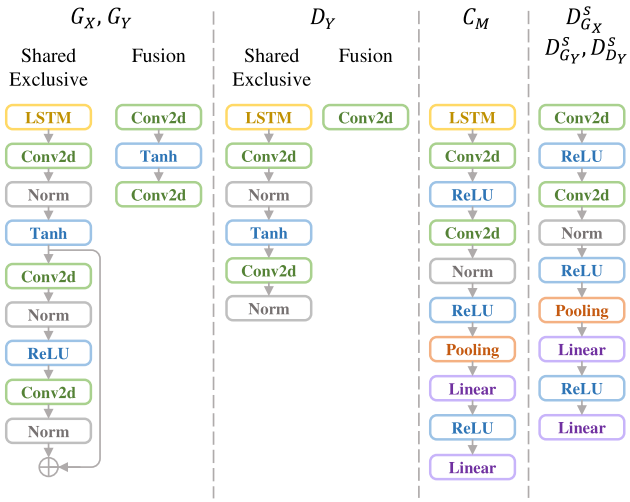



Fig. 2. Detailed architecture.

Let $F^s \in \mathbb{R}^{C\times H\times W}$ and $F^e \in \mathbb{R}^{C\times H\times W}$ be the extracted shared and exclusive features, respectively. These features are concatenated along the channel dimension as follows:

$$F = [F^s, F^e] \in \mathbb{R}^{2C\times H\times W}. \tag{3}$$

To extract channelwise dependencies, both global average pooling (GAP) and global max pooling are utilized

$$F_{\text{avg}} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} F(i,j), \quad F_{\max} = \max_{i,j} F(i,j) \tag{4}$$

where $F_{\text{avg}}, F_{\max} \in \mathbb{R}^{2C\times 1\times 1}$. The pooled features are concatenated

$$F_{\text{pool}} = [F_{\text{avg}}, F_{\max}] \in \mathbb{R}^{2C\times 2\times 1}. \tag{5}$$

Diverse receptive fields are achieved by leveraging multiscale convolutions with varying kernel sizes

$$A_i = \sigma(W_i * F_{\text{pool}}) \tag{6}$$

where $*$ denotes the convolution operation and $A_i$ represents the attention weights learned by the $i$th convolutional kernel. The final attention map is obtained

$$A = \sigma\left(W_{\text{att}} * \text{concat}(A_1, A_2, A_3)\right). \tag{7}$$

The refined feature representation is obtained via elementwise multiplication

$$F_{\text{out}} = F \odot A. \tag{8}$$

To enhance the activation of informative channels while suppressing irrelevant ones, an attention regularization loss is introduced

$$\mathcal{L}_{\text{are}} = \frac{1}{2C}\|A\|_2^2. \tag{9}$$

The overall transformation performed by MRCA can be formulated as follows:

$$F_{\text{att}} = \mathcal{T}_{\text{MRCA}}(F^s, F^e) \tag{10}$$

where $\mathcal{T}_{\text{MRCA}}(\cdot)$ represents the feature refinement function of MRCA.

Accordingly, MRCA leverages multiscale convolutions to capture rich hierarchical dependencies. In addition, the regularization further refines feature selection, ensuring that only the most discriminative channels contribute to fault detection.

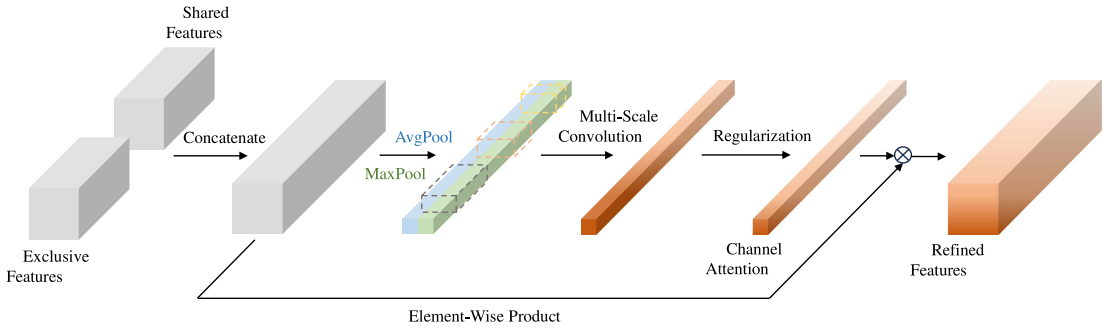



Fig. 3. Multiscale regularized channel attention.

### B. Shared and Exclusive Feature Extraction

To extract shared and exclusive features, the generators and discriminator in SERLM are equipped with a shared feature extractor and an exclusive feature extractor. The shared feature extractor learns mode-invariant representations, with a single extractor consistently applied to data from all modes to ensure invariance. In contrast, the exclusive feature extractors are designed to capture mode-specific information, with a distinct extractor assigned to each mode.

For the generator $G_X$, the shared feature extractor and the exclusive feature extractor can be formulated as:

$$F^s_{G_X} = \mathcal{S}_{G_X}(X), \quad F^{e,i}_{G_X} = \mathcal{E}^i_{G_X}(X) \tag{11}$$

where $\mathcal{S}_{G_X}(\cdot)$ and $\mathcal{E}^i_{G_X}(\cdot)$ represent the shared feature extractor and the $i$th exclusive feature extractor in $G_X$ (with $i$ denoting the mode index), respectively. The extracted features are subsequently processed through the MRCA module to derive weighted feature representations

$$F^{\text{att}}_{G_X} = \mathcal{T}_{\text{MRCA}}(F^s_{G_X}, F^{e,i}_{G_X}). \tag{12}$$

The feature fusion module further aggregates these features to generate the final representation

$$Y = \mathcal{F}_{G_X}(F^{att}_{G_X}) \tag{13}$$

where $\mathcal{F}_{G_X}(\cdot)$ represents the feature fusion function for $G_X$.

The overall transformation performed by the generator $G_X$ can be summarized as

$$Y = G_X(X) = \mathcal{F}_{G_X}\left(\mathcal{T}_{\text{MRCA}}(\mathcal{S}_{G_X}(X), \mathcal{E}_{G_X}(X))\right) \tag{14}$$

where $G_X(\cdot)$ represents the transformation function for $G_X$.

### C. Mode Identification and Domain Transformation

The identification of operating modes is achieved through a mode classifier. Given an input $X$, the classifier predicts its mode label

$$\hat{m} = C_M(X) \tag{15}$$

where $C_M(\cdot)$ represents the classification function. The mode classifier is trained to accurately predict the mode of input data using the mode classification loss

$$\mathcal{L}_{cls} = \mathbb{E}_{\hat{X}\sim p_{\text{data}}}\left[-\sum_{i=1}^{k} m_i \log C_M(\hat{X})\right] \tag{16}$$

where $m_i$ is the ground-truth one-hot mode label.

Notably, the mode classifier not only distinguishes operating modes but also guides feature extraction by determining which exclusive feature extractor should be activated, thereby ensuring that mode-specific information is effectively captured and improving the accuracy of fault detection.

Furthermore, the generators $G_X$ and $G_Y$ establish the bidirectional transformation between the source domain and the feature domain, which is enforced by the cycle consistency loss

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{\hat{X}\sim p_{\text{data}}}\left[\|G_Y(G_X(\hat{X})) - \hat{X}\|_1\right]. \tag{17}$$

### D. Domain and Shared Feature Adversarial Learning

Domain adversarial learning ensures that the features generated by $G_X$ match the distribution of features extracted by FENet, achieved via the domain discriminator $D_Y$. The domain adversarial loss is defined as

$$\begin{aligned}\mathcal{L}_{\text{dad}} = {} & \mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[\log D_Y(\hat{Y})\right] \\ & + \mathbb{E}_{\hat{X}\sim p_{\text{data}}}\left[\log(1 - D_Y(G_X(\hat{X})))\right].\end{aligned} \tag{18}$$

Shared feature adversarial learning enforces the consistency of extracted shared features across modes via the discriminators $D^s_{G_X}$, $D^s_{G_Y}$, and $D^s_{D_Y}$. The corresponding shared feature adversarial loss is formulated as

$$\begin{aligned}\mathcal{L}_{sfa} = {} & \mathbb{E}_{\hat{X}\sim p_{\text{data}}}\left[-\sum_{i=1}^{k} m_i \log D^s_{G_X}(F^s_{G_X})\right] \\ & + \mathbb{E}_{\hat{X}\sim p_{\text{data}}}\left[-\sum_{i=1}^{k} \frac{1}{k} \log D^s_{G_X}(\mathcal{S}_{G_X}(\hat{X}))\right] \\ & + \mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[-\sum_{i=1}^{k} m_i \log D^s_{G_Y}(F^s_{G_Y})\right] \\ & + \mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[-\sum_{i=1}^{k} \frac{1}{k} \log D^s_{G_Y}(\mathcal{S}_{G_Y}(\hat{Y}))\right]\end{aligned}$$

$$+\mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[-\sum_{i=1}^{k} m_i \log D^s_{D_Y}(F^s_{D_Y})\right]$$
$$+\mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[-\sum_{i=1}^{k} \frac{1}{k} \log D^s_{D_Y}(\mathcal{S}_{D_Y}(\hat{Y}))\right]. \quad (19)$$

### E. Feature Regularization and Feature Orthogonalization

To enhance the robustness of shared and exclusive features, a feature regularization term is introduced to encourage sparsity in the learned representations. The feature regularization loss is defined as

$$\mathcal{L}_{\text{fre}} = \mathbb{E}_{\hat{X}\sim p_{\text{data}}}\left[\frac{1}{d}\|F^s_{G_X}\|_1 + \frac{1}{d}\|F^e_{G_X}\|_1\right]$$
$$+\mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[\frac{1}{d}\|F^s_{G_Y}\|_1 + \frac{1}{d}\|F^e_{G_Y}\|_1\right]$$
$$+\mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[\frac{1}{d}\|F^s_{D_Y}\|_1 + \frac{1}{d}\|F^e_{D_Y}\|_1\right] \quad (20)$$

where $d$ denotes the dimensionality of the shared and exclusive features.

Notably, Proposition 1 analyzes the theoretical effect of $\mathcal{L}_{\text{fre}}$, showing that it suppresses redundant feature dimensions and thus enhances sparsity and robustness.

Moreover, a feature orthogonality loss is introduced to enforce orthogonality between the representational spaces of shared and exclusive features. This constraint not only reduces redundancy and enhances complementarity, but also establishes the independence assumption between shared and exclusive features ($F^s \perp F^e$), which is utilized in the theoretical analysis of Theorem 1

$$\mathcal{L}_{\text{ort}} = \mathbb{E}_{\hat{X}\sim p_{\text{data}}}\left[\frac{1}{d}\left\|(F^s_{G_X})^\top F^e_{G_X}\right\|_1\right]$$
$$+\mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[\frac{1}{d}\left\|(F^s_{G_Y})^\top F^e_{G_Y}\right\|_1\right]$$
$$+\mathbb{E}_{\hat{Y}\sim p_{\text{data}}}\left[\frac{1}{d}\left\|(F^s_{D_Y})^\top F^e_{D_Y}\right\|_1\right]. \quad (21)$$

*Proposition 1:* Suppose the unregularized estimate $Z_i^F$ of each feature dimension $F_i$ follows a Gaussian distribution $Z_i^F \sim \mathcal{N}(0, \sigma_i^2)$. Then, the probability that $F_i$ is shrunk to zero by $\mathcal{L}_{fre}$ is $\mathbb{P}(F_i = 0) = 2\Phi(\frac{1}{d\cdot\sigma_i}) - 1$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution.

*Proof:* Let $F_i$ be the $i$th dimension of a shared or exclusive feature vector constrained by the feature regularization loss

$$\mathcal{L}_{\text{fre}} = \frac{1}{d}\sum_{i=1}^{d}|F_i|. \quad (22)$$

To assess the impact of $\mathcal{L}_{\text{fre}}$, a single feature dimension $F_i$ is considered in isolation, leading to the following 1-D optimization objective

$$\min_{F_i}\ \mathcal{L}(F_i) + \frac{1}{d}|F_i| \quad (23)$$

where $\mathcal{L}(F_i)$ denotes the task-related loss associated with $F_i$, excluding the regularization term.

To enable analytical tractability, the unregularized objective $\mathcal{L}(F_i)$ is locally approximated by a second-order Taylor expansion around its minimizer, denoted by $Z_i$

$$\mathcal{L}(F_i) \approx \frac{1}{2}(F_i - Z_i^F)^2 \quad (24)$$

where $Z_i^F$ represents the optimal estimate of the $i$th feature dimension in the absence of $\mathcal{L}_{\text{fre}}$. Due to the inherent stochasticity in model training, $Z_i^F$ can be modeled as a zero-mean Gaussian random variable

$$Z_i^F \sim \mathcal{N}(0, \sigma_i^2) \quad (25)$$

where $\sigma_i$ quantifies the uncertainty or redundancy associated with the corresponding feature component.

Accordingly, the overall objective can be reformulated as

$$\mathcal{J}_F(F_i) = \frac{1}{2}(F_i - Z_i^F)^2 + \frac{1}{d}|F_i| \quad (26)$$

whose closed-form minimizer is

$$F_i^* = \text{sign}(Z_i^F)\cdot\max\left(|Z_i^F| - \frac{1}{d}, 0\right). \quad (27)$$

It follows that $F_i^* = 0$ if and only if $|Z_i^F| \le \frac{1}{d}$. Then

$$\mathbb{P}(F_i = 0) = \mathbb{P}\left(|Z_i^F| \le \frac{1}{d}\right). \quad (28)$$

Given $Z_i^F \sim \mathcal{N}(0, \sigma_i^2)$, we obtain

$$\mathbb{P}(F_i = 0) = \int_{-\frac{1}{d}}^{\frac{1}{d}} \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{z^2}{2\sigma_i^2}\right)\mathrm{d}z$$
$$= 2\Phi\left(\frac{1}{d\cdot\sigma_i}\right) - 1. \quad (29)$$

Notably, a redundant feature dimension $F_i$ tends to exhibit lower variability, reflected by a smaller $\sigma_i$, and is thus more likely to be shrunk to zero. □

*Proposition 2:* The attention regularization loss $\mathcal{L}_{\text{are}}$ reduces the magnitude of attention weights without inducing sparsity. In particular, the variance of the regularized attention satisfies: $\mathbb{E}[A_i^{*2}] = (\frac{C}{C+1})^2\mathbb{E}[(Z_i^A)^2]$.

*Proof:* As defined in (9), the attention regularization loss is given by

$$\mathcal{L}_{\text{are}} = \frac{1}{2C}\sum_{i=1}^{2C} A_i^2. \quad (30)$$

Analogous to Theorem 1, suppose the unregularized estimate $Z_i^A$ of the $i$th attention weight $A_i$ follows a Gaussian distribution, i.e., $Z_i^A \sim \mathcal{N}(0, \tau_i^2)$. The local objective can be expressed as

$$\mathcal{J}_A(A_i) = \frac{1}{2}(A_i - Z_i^A)^2 + \frac{1}{2C}A_i^2 \quad (31)$$

whose closed-form minimizer is given by

$$A_i^* = \frac{C}{C+1}Z_i^A. \quad (32)$$

This implies that channel attention is linearly scaled rather than set to zero. Moreover, the variance of the regularized attention satisfies

$$\mathbb{E}[A_i^{*2}] = \left(\frac{C}{C+1}\right)^2 \tau_i^2 < \tau_i^2 = \mathbb{E}[(Z_i^A)^2]. \tag{33}$$

Hence, $\mathcal{L}_{\text{are}}$ encourages stable and generalizable attention allocation through smooth energy attenuation, without explicitly enforcing sparsity. ■

### F. Overall Optimization Objective

Based on (9) and (16)–(21), the total loss function is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cyc}} + \mathcal{L}_{\text{dad}} + \mathcal{L}_{sfa} + \mathcal{L}_{\text{fre}} + \mathcal{L}_{\text{are}} + \mathcal{L}_{\text{ort}}. \tag{34}$$

Consequently, the overall optimization objective is formulated as

$$\theta^* = \arg \min_{\substack{C, \\ G_X, G_Y}} \max_{\substack{D_Y, D^s_{G_X}, \\ D^s_{G_Y}, D^s_{D_Y}}} \mathcal{L}_{\text{total}} \tag{35}$$

where $\theta$ represents the parameters of SERLM.

### G. Monitoring Statistics

In this section, two statistics, $T_s$ and $T_r$, are introduced. $T_s$ is constructed from shared feature representations to identify global anomalies across different modes, ensuring consistency in detecting cross-mode variations. Meanwhile, $T_r$ is based on reconstruction error, which captures localized anomalies specific to individual modes. The combination of these two statistics enhances the sensitivity and robustness of fault detection in multimode processes.

*1) $T_s$:* Let $F^s$ denote the shared feature representation. A GAP operation is applied to $F^s$

$$f^s = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} F^s(i,j). \tag{36}$$

The mean and covariance of the shared feature distribution are estimated as

$$\mu_s = \frac{1}{N} \sum_{i=1}^{N} f_i^s, \quad \Sigma_s = \frac{1}{N-1} \sum_{i=1}^{N} (f_i^s - \mu_s)(f_i^s - \mu_s)^T. \tag{37}$$

Subsequently, $T_s$ is formulated as follows:

$$T_s = (f^s - \mu_s)^T \Sigma_s^{-1} (f^s - \mu_s). \tag{38}$$

*2) $T_r$:* $T_r$ evaluates the deviation between the input sample and its reconstruction by the generators

$$T_r = \|G_Y(G_X(\hat{X})) - \hat{X}\|_2^2. \tag{39}$$

### H. Monitoring Strategy

*1) Data Collection:* Let $\mathcal{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(k)}\}$ denote the training data collected from $k$ distinct operating modes, and $\mathcal{Y} = \{Y^{(1)}, Y^{(2)}, \ldots, Y^{(k)}\}$ represent the corresponding features generated by FENet. $X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}] \in \mathbb{R}^{n_i \times m}$ represents the data collected from the $i$th mode, $n_i$ is the number of samples in mode $i$, and $m$ denotes the dimensionality of each sample.

**Algorithm 1:** Process Monitoring Strategy for SERLM.

**Offline Training**

1: Collect multimode training data $\mathcal{X}$.
2: Extract intrinsic features $\mathcal{Y}$ from $\mathcal{X}$ using FENet.
3: Standardize $\mathcal{X}$ using (41) to obtain $\hat{\mathcal{X}}$. Similarly, derive $\hat{\mathcal{Y}}$ from $\mathcal{Y}$.
4: Randomly initialize the parameters of SERLM.
5: **for** training epoch from 1 to $N_1$ **do**
6: Input training data $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$.
7: Compute the loss.
8: Optimize the model parameters.
9: **end for**
10: Calculate the statistics $T_s$ and $T_r$ for each training sample using (38) and (39), respectively.
11: Estimate control limits $s_s^{\lim}$ and $s_r^{\lim}$ using KDE.

**Online Monitoring**

12: Standardize and apply sliding window processing to the arriving sample $x_{n+1}$.
13: Identify the mode of $x_{n+1}$ via (15).
14: Compute the statistics $T_{s,n+1}$ and $T_{r,n+1}$ via (38) and (39), respectively.
15: Determine the state of $x_{n+1}$ according to (42).

*2) Data Preprocessing:* The global mean and standard deviation across all modes are computed as

$$\mu = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_j^{(i)}, \quad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_j^{(i)} - \mu)^2} \tag{40}$$

where $N = \sum_{i=1}^{k} n_i$ is the total number of samples. Then, $X^{(i)}$ can be standardized as

$$\hat{X}^{(i)} = (X^{(i)} - \mu)/\sigma. \tag{41}$$

Subsequently, $\hat{\mathcal{X}} = \{\hat{X}^{(1)}, \hat{X}^{(2)}, \ldots, \hat{X}^{(k)}\}$ is obtained.

Similarly, $\mathcal{Y}$ is preprocessed to generate $\hat{\mathcal{Y}}$.

*3) Control Limit Calculation:* Kernel density estimation (KDE) [19] is utilized to estimate the control limits $s_s^{\lim}$ and $s_r^{\lim}$ for the statistics $T_s$ and $T_r$, respectively, based on a confidence level of $\alpha = 0.99$.

*4) Decision Rule:* During the online monitoring phase, the decision rule for the incoming sample $x_{n+1}$ is defined as

$$\begin{cases} x_{n+1} \text{ is normal}, & \text{if } T_{s,n+1} < s_s^{\lim} \text{ and } T_{r,n+1} < s_r^{\lim} \\ x_{n+1} \text{ is faulty}, & \text{otherwise.} \end{cases} \tag{42}$$

The details of the monitoring strategy are comprehensively outlined in Algorithm 1.

*Theorem 1:* Suppose that the data from the $i$th mode can be modeled as $X^{(i)} = g_s(F^s) + g_e^{(i)}(F^{e,i})$, where $F^s \in \mathbb{R}^{n_i \times d_s}$ denotes the shared feature, $F^{e,i} \in \mathbb{R}^{n_i \times d_s}$ represents the exclusive feature corresponding to the $i$th mode, and $g_s$ and $g_e^{(i)}$ denote the functions that decode shared and exclusive features, respectively. Then, the shared representation $F^s$ is identifiable

up to an affine orthogonal transformation, i.e., $F'^s = UF^s + b$, where $UU^\top = I_{d_s}$ and $b \in \mathbb{R}^{d_s}$.

*Proof:* Suppose that $X^{(i)}$ can be expressed by two different decompositions

$$X^{(i)} = g_s(F^s) + g_e^{(i)}(F^{e,i}) \tag{43}$$

$$X^{(i)} = g'_s(F'^s) + g'^{(i)}_e(F'^{e,i}). \tag{44}$$

Equating the two decompositions yields

$$g_s(F^s) + g_e^{(i)}(F^{e,i}) = g'_s(F'^s) + g'^{(i)}_e(F'^{e,i}). \tag{45}$$

Differentiate both sides of (45) with respect to $F^s$, then it follows that:

$$\frac{\partial}{\partial F^s}\left(g_s(F^s) + g_e^{(i)}(F^{e,i})\right) = \frac{\partial}{\partial F^s}\left(g'_s(F'^s) + g'^{(i)}_e(F'^{e,i})\right). \tag{46}$$

Consequently, it holds that

$$\begin{aligned}\frac{\partial g_s(F^s)}{\partial F^s} &= \frac{\partial g'_s(F'^s)}{\partial F^s} + \frac{\partial g'^{(i)}_e(F'^{e,i})}{\partial F^s} - \frac{\partial g_e^{(i)}(F^{e,i})}{\partial F^s} \\ &= \frac{\partial g'_s(F'^s)}{\partial F^s} + \frac{\partial g'^{(i)}_e(F'^{e,i})}{\partial F^s} - \frac{\partial g_e^{(i)}(F^{e,i})}{\partial F^{e,i}} \cdot \frac{\partial F^{e,i}}{\partial F^s}.\end{aligned} \tag{47}$$

According to (21), $F^{e,i}$ and $F^s$ are assumed to be independent, i.e., $F^{e,i} \perp F^s$, then we have $\frac{\partial F^{e,i}}{\partial F^s} = 0$. Therefore, (46) can be rewritten as

$$\frac{\partial g_s(F^s)}{\partial F^s} = \frac{\partial g'_s(F'^s)}{\partial F^s} + \frac{\partial g'^{(i)}_e(F'^{e,i})}{\partial F^s}. \tag{48}$$

This leads to

$$\frac{\partial g'^{(i)}_e(F'^{e,i})}{\partial F^s} = \frac{\partial g_s(F^s)}{\partial F^s} - \frac{\partial g'_s(F'^s)}{\partial F^s}. \tag{49}$$

It is worth noting that (49) holds across all modes, with $F^s$, $F'^s$, $g_s$, and $g'_s$ remaining consistent for every mode. Thus, the following identity holds:

$$\frac{\partial g'^{(1)}_e(F'^e_1)}{\partial F^s} = \frac{\partial g'^{(2)}_e(F'^e_2)}{\partial F^s} = \cdots = \frac{\partial g'^{(k)}_e(F'^e_k)}{\partial F^s}. \tag{50}$$

As a result

$$\frac{\partial g'^{(i)}_e(F'^{e,i})}{\partial F^s} = \frac{\partial g'^{(i)}_e(F'^{e,i})}{\partial F'^{e,i}} \cdot \frac{\partial F'^{e,i}}{\partial F^s} = \mathbf{C}, (i = 1, 2, \ldots, k) \tag{51}$$

where $\mathbf{C}$ denotes a constant matrix independent of $i$. If $\mathbf{C} \neq \mathbf{0}$, then $F'^e_1, \ldots,$ and $F'^e_k$ would share common features associated with $F^s$. However, according to (17), (19), and (21), $F'^e_1, \ldots,$ and $F'^e_k$ represent exclusive features with no overlap, leading to a contradiction. Hence, $\mathbf{C} = \mathbf{0}$. Accordingly, (48) can be reformulated as

$$\frac{\partial g_s(F^s)}{\partial F^s} = \frac{\partial g'_s(F'^s)}{\partial F^s}. \tag{52}$$

Noting that

$$\frac{\partial g'_s(F'^s)}{\partial F^s} = \frac{\partial g'_s(F'^s)}{\partial F'^s} \cdot \frac{\partial F'^s}{\partial F^s} \tag{53}$$

it follows from (52) and (53) that

$$J_s = J'_s \cdot \frac{\partial F'^s}{\partial F^s} \tag{54}$$

where $J_s = \frac{\partial g_s(F^s)}{\partial F^s}$ and $J'_s = \frac{\partial g'_s(F'^s)}{\partial F'^s}$. It is worth noting that $g_s$ and $g'_s$ are assumed to be continuously differentiable and locally invertible mappings. Then, we obtain

$$\operatorname{rank}(J_s) = \operatorname{rank}(J'_s) = d_s \tag{55}$$

which means that $J_s$ and $J'_s$ are full rank. It follows from (54) and (55) that $\frac{\partial F'^s}{\partial F^s}$ is invertible. To enhance the structural identifiability of the affine transformation, the following is introduced:

$$\left(\frac{\partial F'^s}{\partial F^s}\right)\left(\frac{\partial F'^s}{\partial F^s}\right)^\top = I_{d_s} \tag{56}$$

where $I_{d_s}$ denotes the identity matrix of size $d_s \times d_s$. Without loss of generality, let $\frac{\partial F'^s}{\partial F^s} = U$, $U \in \mathbb{R}^{d_s \times d_s}$, it follows from (56) that

$$UU^\top = I_{d_s} \tag{57}$$

$$F'^s = UF^s + b \tag{58}$$

where $b \in \mathbb{R}^{d_s}$.

Therefore, $F^s$ is identifiable up to an affine orthogonal transformation. □

Notably, the identifiability guarantee established in Theorem 1 is concretely achieved in the SERLM architecture through three design choices. First, a single shared feature extractor for all modes and distinct exclusive feature extractors for each mode capture mode-invariant features and mode-specific variations, respectively. Second, shared feature discriminators are introduced to enforce consistency of the shared features across modes, thereby aligning them with the theoretical assumption of invariance. Third, the feature orthogonality loss encourages independence between shared and exclusive features, which corresponds to the independence condition required in the proof.

### I. Performance Evaluation Metrics

The detection performance is quantitatively evaluated using two metrics: the false alarm rate (FAR) and the fault detection rate (FDR), which are defined as follows:

$$\text{FAR} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TN}}} \times 100\% \tag{59}$$

$$\text{FDR} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \times 100\% \tag{60}$$

where $N_{\text{FP}}$, $N_{\text{TN}}$, $N_{\text{TP}}$, and $N_{\text{FN}}$ denote the numbers of false positive, true negative, true positive, and false negative samples, respectively.

## IV. Experimental Validation

This section presents the experimental platform and validates the effectiveness and superior performance of the proposed SERLM in comparison with recursive PCA (RPCA) [8],

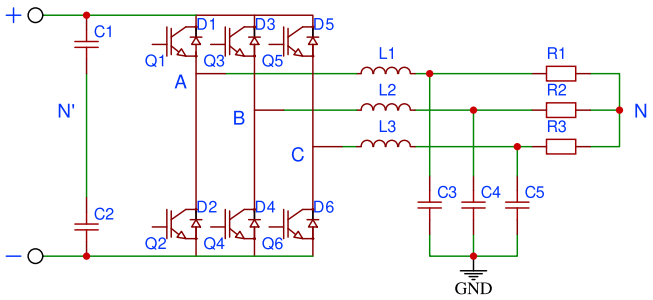



Fig. 4. Circuit schematic diagram.

MPPCA [11], finite Gaussian mixture model (FGMM) [20], PSAT [10], DSRN [12], and ablation studies.

### A. Experimental Platform

To evaluate the effectiveness of the proposed method in detecting incipient faults in multimode inverter systems, an experimental platform was developed. The system is implemented based on a three-phase full-bridge inverter topology, where six insulated gate bipolar transistors (IGBTs) function as the primary switching devices for efficient power conversion. The schematic diagram of the inverter system is depicted in Fig. 4, while Fig. 5 illustrates the practical experimental platform.

The experimental platform comprises three core components: the inverter module, a host computer, and a dedicated control unit. The inverter module converts dc power into a three-phase ac output through the coordinated switching of IGBTs. The host computer runs control algorithms, processes acquired data, and facilitates communication with the control unit. The control unit is responsible for generating real-time switching signals and implementing closed-loop regulation to ensure the stable operation of the inverter system.

### B. Data Acquisition

To effectively characterize the system state, three-phase voltage waveforms are acquired from the load terminals, and six statistical features are extracted for each phase: frequency, mean, root mean square, peak-to-peak value, clearance factor, and kurtosis. Consequently, each sample consists of 18 variables, offering a comprehensive representation of the voltage signals.

To account for varying operating modes, data acquisition is performed under three different load impedance settings: 20, 50, and 65 Ω. These variations are introduced to simulate distinct operating scenarios, thereby facilitating the validation of the fault detection method's robustness. Each mode includes 8000 samples, providing a sufficient dataset for assessing system behavior.

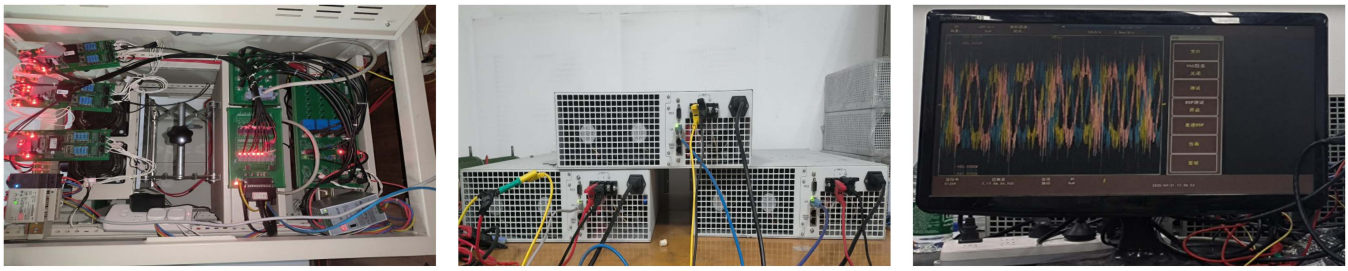
Fault conditions are introduced by connecting a parallel resistor across the inductor, thereby altering the system's dynamic characteristics. As shown in Table II, the parallel impedance varies across modes to simulate different fault severities, enabling a thorough evaluation of fault detection performance. In each mode, the fault is initiated at the 6001st sample. The first 4000 samples are allocated for model training, while the remaining 4000 samples are utilized for testing. To facilitate a systematic assessment of fault detection capability across different modes, the test data from all three modes are integrated into a single dataset, where the first 4000 samples correspond to Mode 1, the next 4000 samples to Mode 2, and the final 4000 samples to Mode 3.

### C. Parameter Configuration

In RPCA, the $T^2$ and $Q$ statistics are calculated, with the cumulative percentage of variance set to 0.8. The latent dimensionality of each subspace for MPPCA is fixed at $q = 15$. The number of mixture components in FGMM is set to 3. For SERLM, the model is trained for 100 epochs ($N_1 = 100$) with a learning rate of $\eta = 0.0001$ to ensure stable convergence. In ablation studies, the parameters remain identical to those in SERLM to isolate the impact of structural variations, thereby enabling fair performance comparisons.

### D. Experimental Results and Analysis

The monitoring performance of various methods is illustrated in Fig. 6, while detailed quantitative results in terms of FARs and FDRs are reported in Table I . Among these methods, MPPCA, which extends PCA to multimode cases, exhibits limited capability in capturing intrinsic features of incipient faults due to its strong Gaussian mixture assumption, achieving FDRs of only 7.27% and 8.38% based on the $T^2$ and $Q$ statistics, respectively. RPCA, constrained by its linear subspace projection and limited adaptability, struggles to handle multimode processes with large variations across modes. As a result, it only detects a portion of faults, with an FDR of 54.62% for the $T^2$ statistic, but this comes at the cost of a relatively high FAR of 14.53% . FGMM demonstrates comparatively better monitoring performance by allowing flexible mixture modeling, attaining an FDR of 50.23% and an FAR of 4.68% ; however, its sensitivity to initialization and tendency to overfit limit its robustness, leaving substantial room for improvement. PSAT and DSRN achieve relatively low FARs (5.98% and 5.32% , respectively), indicating their ability to control false alarms. However, their FDRs (86.67% and 92.33% ) are notably lower than that of SERLM, revealing insufficient sensitivity to incipient faults. SERLM exhibits superior detection performance, achieving FDRs of 96.98% and 98.58% based on $T_r$ and $T_s$ statistics, respectively, while maintaining low FARs. The advantageous performance of SERLM is primarily attributed to the integration of FENet, which facilitates the representation of incipient faults, and MRCA, which effectively fuses shared and exclusive features. To further validate the contribution of FENet and MRCA, ablation studies are conducted. Results indicate that SERLM w/o FENet (removing FENet from SERLM) leads to a significant degradation in performance, with FDRs of $T_r$ and $T_s$ dropping to 75.07% and 88.63% , respectively. Moreover, when MRCA is excluded from the framework (SERLM w/o MRCA), the FARs increase to 7.73% and 6.20% for $T_r$ and $T_s$, respectively. These results demonstrate that both FENet and MRCA are essential components for maintaining the overall effectiveness of SERLM.



(a) (b) (c)

Fig. 5. Practical experimental platform. (a) The practical circuit. (b) The electronic loads. (c) The power analyzer.

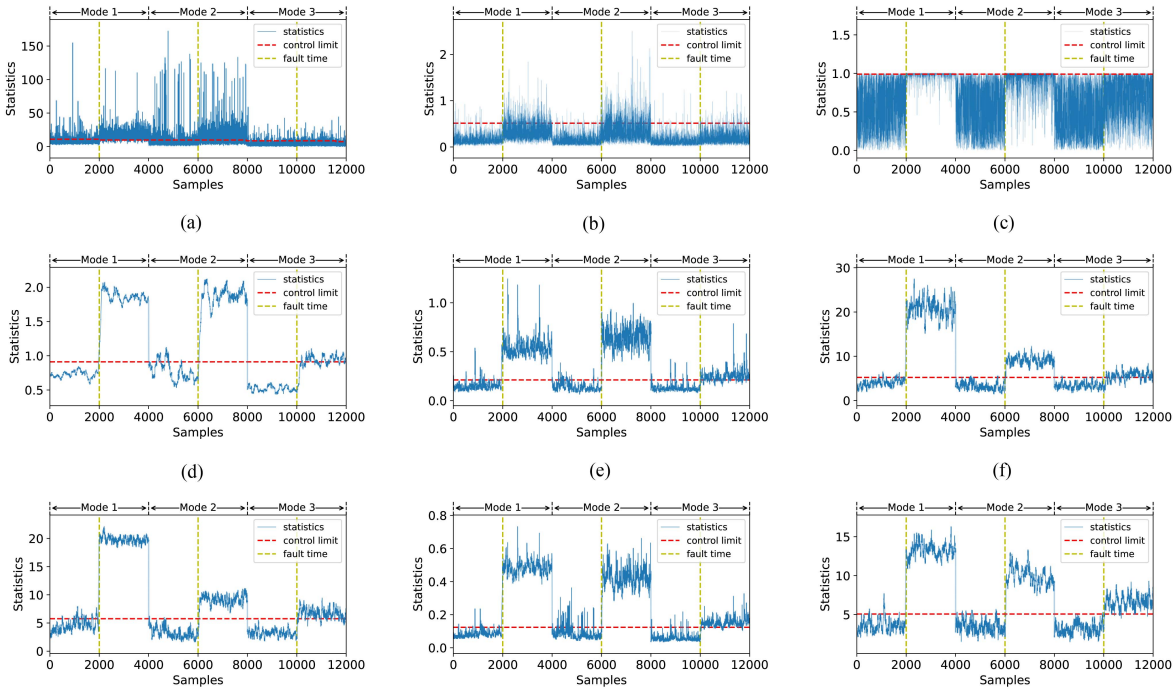



Fig. 6. Performance comparison in the inverter system. (a) RPCA $(T^2)$. (b) MPPCA $(Q)$. (c) FGMM. (d) PSAT. (e) DSRN. (f) SERLM w/o FENet $(T_s)$. (g) SERLM w/o MRCA $(T_s)$. (h) SERLM $(T_r)$. (i) SERLM $(T_s)$.

TABLE I
FARS AND FDRS IN THE INVERTER SYSTEM

| Methods | RPCA | | MPPCA | | FGMM | PSAT | DSRN | SERLM w/o FENet | | SERLM w/o MRCA | | SERLM | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T^2$ | $Q$ | $T^2$ | $Q$ | | | | $T_r$ | $T_s$ | $T_r$ | $T_s$ | $T_r$ | $T_s$ |
| FAR(%) | 14.53 | 15.50 | 1.27 | 1.42 | 4.68 | 5.98 | 5.32 | 3.38 | 3.45 | 7.73 | 6.20 | 2.23 | 2.70 |
| FDR(%) | 54.62 | 43.47 | 7.27 | 8.38 | 50.23 | 86.67 | 92.33 | 75.07 | 88.63 | 97.30 | 94.68 | 96.98 | 98.58 |

In addition, to visually demonstrate the superior feature extraction capability of SERLM, t-distributed stochastic neighbor embedding [21] is employed to project both the raw data and the extracted shared features into a 2-D space. As illustrated in Fig. 7, the normal and faulty samples from different modes exhibit significant overlap in the raw data, making them difficult to distinguish. While SERLM w/o FENet and SERLM w/o MRCA exhibit improved clustering in the shared feature space, the class boundaries remain ambiguous. In contrast, the shared features extracted by SERLM form compact and clearly separable clusters for both normal and faulty samples, highlighting their mode-invariance and representational effectiveness.

Moreover, we quantified SERLM's computational complexity and inference latency and compared them against PSAT and DSRN, as summarized in Table III. SERLM exhibits the most favorable efficiency profile, requiring fewer parameters

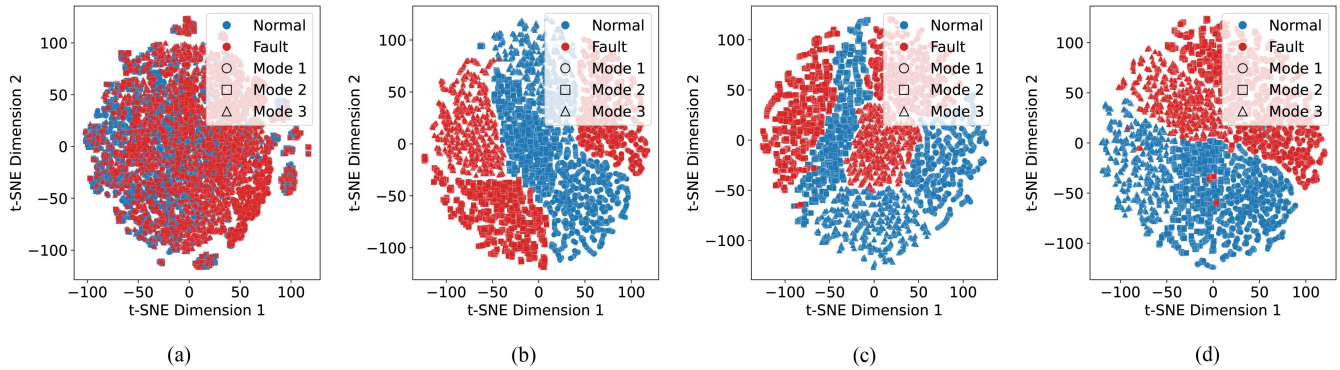



Fig. 7. Visualization of data. (a) Raw data. (b) SERLM w/o FENet. (c) SERLM w/o MRCA. (d) SERLM.

TABLE II
FAULT CONDITION

| Mode | Load Impedance (Ω) | Parallel Impedance (Ω) |
|---|---|---|
| 1 | (20, 20, 20) | 100 |
| 2 | (50, 50, 50) | 200 |
| 3 | (65, 65, 65) | 700 |

TABLE III
COMPARISON OF MODEL COMPLEXITY AND INFERENCE LATENCY (BATCH = 1)

| Method | Params (M) | FLOPs (G) | Latency (s/sample) | FPS |
|---|---|---|---|---|
| PSAT | 2.89 | 1.36 | 0.0090 | 110.54 |
| DSRN | 3.05 | 1.64 | 0.0091 | 109.47 |
| SERLM | 2.72 | 1.04 | 0.0087 | 114.29 |

(2.72 M) and lower floating-point operation (1.04 G) than PSAT (2.89 M/1.36 G) and DSRN (3.05 M/1.64 G), while delivering the highest throughput. With a per-sample latency of 8.7 ms (114.29 FPS), SERLM satisfies real-time online monitoring requirements typical of industrial inverter processes.

## V. CONCLUSION

In this article, a novel SERLM was proposed for incipient fault detection in multimode power inverter systems. SERLM extracted shared features across all modes and exclusive features for each individual mode. In addition, shared feature discriminators were introduced to enforce the invariance of shared features across modes, while a feature orthogonality loss ensured the independence between shared and exclusive features. These features are further integrated through MRCA to enhance the representation of incipient faults. Moreover, the feature extraction capability of SERLM is rigorously analyzed from a theoretical perspective. Finally, a power inverter system was utilized to validate the effectiveness of SERLM, and experimental results demonstrated that SERLM achieves superior detection performance, which is primarily attributed to the integration of FENet for capturing incipient fault representations and the effective fusion of shared and exclusive features through MRCA. Notably, SERLM requires complete data from all modes during training and necessitates retraining upon the emergence of new modes. Consequently, future work will explore incorporating continual learning strategies to enhance adaptability to previously unencountered modes.

## REFERENCES

[1] K. Zeb et al., "A comprehensive review on inverter topologies and control strategies for grid connected photovoltaic system," *Renew. Sustain. Energy Rev.*, vol. 94, pp. 1120–1141, 2018.
[2] M. Wang, D. Zhou, and M. Chen, "Hybrid variable monitoring: An unsupervised process monitoring framework with binary and continuous variables," *Automatica*, vol. 147, 2023, Art. no. 110670.
[3] Y. Yang and P. Zhang, "A novel bond wire fault detection method for IGBT modules based on turn-on gate voltage overshoot," *IEEE Trans. Power Electron.*, vol. 36, no. 7, pp. 7501–7512, Jul. 2020.
[4] M. Wang, F. Cheng, M. Xie, G. Qiu, and J. Zhang, "Intensive multi-order feature extraction for incipient fault detection of inverter system," *IEEE Trans. Power Electron.*, vol. 40, no. 2, pp. 3543–3552, Feb. 2024.
[5] X. Gu, Y. Liu, G. Zhang, Y. Cao, and C. Xia, "Multi-mode carrier-based PWM strategy of open-end winding PMSM with isolated DC sources," *IEEE Trans. Power Electron.*, vol. 40, no. 1, pp. 1826–1835, Jan. 2024.
[6] Y. Li, W. Sun, J. Liu, Y. Liu, J. Hu, and Z. He, "Multimode control of WPT systems for efficiency improvement in wide output voltage range applications," *IEEE Trans. Power Electron.*, vol. 38, no. 12, pp. 14818–14829, Dec. 2023.
[7] D. Serrano, R. Ramos, P. Alou, J. A. Oliver, and J. A. Cobos, "Multimode modulation with ZVS for a single-phase single-stage inverter," *IEEE Trans. Power Electron.*, vol. 35, no. 5, pp. 5319–5330, May 2019.
[8] W. Li, H. H. Yue, S. Valle-Cervantes, and S. J. Qin, "Recursive PCA for adaptive process monitoring," *J. Process Control*, vol. 10, no. 5, pp. 471–486, 2000.
[9] K. Zhang, K. Peng, and J. Dong, "A common and individual feature extraction-based multimode process monitoring method with application to the finishing mill process," *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4841–4850, Nov. 2018.
[10] Y. Wang, Y. He, B. Kang, J. Liu, and C. Sun, "ProbSparse attention-based fault diagnosis for industrial robots under different working conditions," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024.
[11] R. Sharifi and R. Langari, "Nonlinear sensor fault diagnosis using mixture of probabilistic PCA models," *Mech. Syst. Signal Process.*, vol. 85, pp. 638–650, 2017.
[12] M. Xu, Y. Han, X. Ding, H. Shao, and Y. Shao, "Decision self-regulating network for imbalanced working conditions identification in the application of gearbox intelligent fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.
[13] C. Cheng, Y. Ju, S. Xu, Y. Lv, and H. Chen, "Local linear generalized autoencoder-based incipient fault detection for electrical drive systems of high-speed trains," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12422–12430, Nov. 2023.
[14] C. Wang, K. Sheng, Z. Liu, J. Wang, and M. Wang, "Incipient fault detection based on multi-scale time series feature extraction," *IEEE Trans. Instrum. Meas.*, vol. 74, 2025, Art. no. 3524010.
[15] J. Shang, M. Chen, H. Ji, and D. Zhou, "Recursive transformed component statistical analysis for incipient fault detection," *Automatica*, vol. 80, pp. 313–327, 2017.

[16] D. Liu, M. Wang, and M. Chen, "Feature ensemble net: A deep framework for detecting incipient faults in dynamical processes," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 8618–8628, Dec. 2022.
[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2223–2232.
[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11534–11542.
[19] P. Phaladiganon, S. B. Kim, V. C. P. Chen, and W. Jiang, "Principal component analysis-based control charts for multivariate nonnormal distributions," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 3044–3054, 2013.
[20] J. Yu and S. J. Qin, "Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models," *AIChE J.*, vol. 54, no. 7, pp. 1811–1829, 2008.
[21] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

**Shuihui Rao** (Graduate Student Member, IEEE) received the B.Eng. degree in automation from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2023. He is currently working toward the master's degree in control science and engineering with the Department of Automation, Tsinghua University, Beijing.

His research interests include process monitoring, fault detection, deep learning, and ensemble learning.

**Yu Shuai** (Graduate Student Member, IEEE) received the B.Eng. degree in measurement and instrument in 2023 from the School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, Sichuan, China, where he is currently working toward the M.Eng. degree in instrument science and technology.

His research focuses on anomaly detection based on deep learning.

**Zhihua Xiong** (Member, IEEE) received the B.S. and M.S. degrees in automatic control from Northeastern University, Shenyang, China, in 1993 and 1996, respectively, and the Ph.D. degree in control theory and control engineering from Tsinghua University, Beijing, China, in 2000.

From 2000 to 2003, he was a Postdoctoral Research Associate with the Centre for Process Analytics and Control Technology, University of Newcastle, Newcastle upon Tyne, U.K. He is currently an Associated Professor with the Department of Automation, Tsinghua University. His research interests include anomaly detection, fault diagnosis, intelligent control, and neural networks.

**Min Wang** (Member, IEEE) received the B.Eng. degree in measurement and control technology and instrument from the School of Information Engineering, Chang'an University, Xi'an, China, in 2015, the M.Eng. degree in industrial engineering from the School of Reliability and Systems Engineering, Beihang University, Beijing, China, in 2018, and the Ph.D. degree in control science and engineering from the Department of Automation, Tsinghua University, Beijing, in 2022.

He was a Postdoctoral Researcher with the Department of Advanced Design and Systems Engineering, City University of Hong Kong, Hong Kong, and the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China. He was an Assistant Professor of the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, where he is currently an Associate Professor. His research interests include anomaly detection, fault diagnosis, life prediction, and health management.