processes

MDPI

Article

Fault Diagnosis of Reciprocating Compressor Crosshead Nut Loosening Using Swin Transformer-Based Ordered Transfer Learning

Zhaozeng Liu [1], Yiyang Geng [1], Xidong Wang [2], Yanhui Chen [1] and Weiwei Xu [1,3,*]

[1] College of New Energy, China University of Petroleum (East China), Qingdao 266580, China; liuzz@upc.edu.cn (Z.L.); 2115010107@s.upc.edu.cn (Y.G.); chenyanhui2022@163.com (Y.C.)
[2] PetroChina Tuha Oilfield Company, Hami 839009, China; wangxdxf@petrochina.com.cn
[3] Key Laboratory of Advanced Nuclear Energy Design and Safety, Ministry of Education, University of South China, Hengyang 421001, China
* Correspondence: xuweiwei@upc.edu.cn; Tel.: +86-152-5322-2583

Abstract

The connecting nut between the piston rod and the crosshead is a key fastener of reciprocating compressors, and its loosening may lead to unplanned shutdowns and even safety accidents. Fault severities are commonly divided into normal, slight, moderate and severe states, but accurate data-driven diagnosis under variable working conditions remains difficult because labeled experimental fault data are scarce and the fault features of different severities are separated in scale. This paper proposes the Swin Transformer Deep Ordered Transfer Network with Multi-Scale Feature Fusion, CORAL Alignment and Ordinal Regression (Swin-MCO). The source domain is a rigid-flexible coupled multibody simulation validated against bench measurements, and the target domain is a test-bench dataset covering 25 speed and pressure conditions, of which only 20% are labeled for training. The model extracts hierarchical features from 224 × 224 angular-domain texture images, aligns second-order statistics across domains and encodes the physical fault progression in the classifier. The diagnosis accuracies reach 95.80%, 95.57% and 96.72% on the three unseen test sets, and misclassifications are confined to adjacent severity levels. The results indicate that the proposed model provides an effective solution for cross-domain diagnosis of connecting nut loosening. Future work will focus on industrial deployment, lightweight edge implementation and uncertainty quantification.

Keywords: multi-scale feature fusion; CORAL alignment; ordinal regression; deep transfer learning; domain adaptation; domain generalization

Check for updates

Academic Editors: Xiaoyu Jiang, Yuanqiang Zhou, Le Yao and Zheren Zhu

Received: 25 October 2026
Revised: 16 November 2026
Accepted: 17 November 2026
Published: 20 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As the core equipment in the petrochemical, coal chemical and natural gas transportation industries, reciprocating compressors feature a wide exhaust pressure range, high thermal efficiency and good adaptability to various media, and have also been widely applied in the nuclear power industry to support safety-critical air and process-gas systems [1,2]. A reciprocating compressor is a coupled mechanical system with multiple degrees of freedom, consisting of crossheads, piston rods, crankshafts, gas valves and sealing assemblies. Its complex motion mechanism makes it prone to various faults under variable operating conditions and continuous high-load industrial operation [3,4]. According to statistics from the European Reciprocating Compressor Forum and the American Petroleum Institute, the

moving mechanism is one of the parts most susceptible to failures. Since it is enclosed inside the unit, its faults are highly concealed and sudden, and may trigger serious safety accidents [5,6]. As a key fastener connecting the piston rod and the crosshead, the crosshead connecting nut is subjected to alternating combined piston forces as well as driving forces transmitted by the crankshaft and connecting rod [7]. Its working condition directly affects the safe operation of the entire unit. Therefore, research on the loosening fault diagnosis of this connecting nut is of great engineering significance. Extensive studies have been devoted to the loosening mechanism and monitoring of threaded connections [8–11], while accurate data-driven diagnosis of the looseness severity of connecting nuts across working conditions remains an open problem.

Deep learning has achieved remarkable progress in mechanical fault diagnosis thanks to its powerful capability for automatic feature extraction. The CNN family has been continuously optimized for local feature extraction, evolving from early convolutional models such as AlexNet [12], to ResNet [13] which addresses the degradation problem of deep networks via residual learning, and further to MobileNet [14] that reduces parameter volume using depthwise separable convolutions. In recent years, the Transformer architecture has been gradually applied to fault diagnosis by virtue of its self-attention mechanism that excels at global information modeling. Swin Transformer [15] introduces hierarchical structures with shifted windows, achieving an excellent balance among local feature extraction, global information integration and computational efficiency. Its self-attention mechanism inherently possesses superior ability to capture long-range features of periodic transient impacts caused by loosening connecting nuts. Transformer-based architectures have also been successfully applied to bearing and bolt looseness diagnosis in recent years [16,17]. Nevertheless, all the aforementioned models rely on the assumption that training and test data follow the same distribution pattern, which limits their application to variable-condition scenarios. Reciprocating compressors in industrial sites usually operate under variable working conditions, and the difficulty in acquiring on-site fault data further limits the generalization ability of pure deep learning models in cross-domain scenarios. In recent years, deep-learning-based compressor diagnosis has continued to advance, including VSG-CNN-based compressor diagnosis [18], spatio-temporal deep belief network diagnosis [19], and improved SSA-VMD-, AO-VMD- and CEEMDAN-based feature extraction [20–22].

To break this constraint, deep transfer learning has become a vital approach for cross-domain fault diagnosis, mainly covering three directions: adversarial-, discrepancy- and parameter-based transfer learning. Adversarial-based methods extract domain-invariant features through the game between feature extractors and domain discriminators, with DANN [23] and CDAN [24] being representative. Discrepancy-based methods align feature distributions via explicit distance metrics, such as the MMD-based DAN [25] and covariance-aligned Deep CORAL [26]. Parameter-based methods transfer knowledge through pre-training and fine-tuning [27], also applied to reciprocating compressor diagnosis [28]. Recent studies have further enriched these directions, including multi-scale adversarial domain adaptation for remaining useful life prediction [29], domain-adversarial speed transfer for scroll compressors [30] and deep transfer metric learning for cross-condition diagnosis [31]. Among them, adversarial-based methods suffer from high computational complexity and training instability, while parameter-based methods rely heavily on pre-training over large source datasets. By contrast, the discrepancy-based Deep CORAL aligns second-order statistics at low computational cost and suits limited labeled data, so it is adopted as the domain adaptation module. Feature alignment alone, however, is insufficient under variable conditions; the MixStyle strategy [32] addresses this by mixing channel statistics to simulate unknown working conditions. On the classification side, an ordinal regression loss [33] is further introduced to enforce order constraints among fault severity

levels, since the loosening fault evolves progressively from normal to severe. The proposed Swin-MCO model integrates these three strategies to address the challenges.

Although the above methods offer effective improvement ideas from the perspectives of domain adaptation, domain generalization and ordinal classification, three coupled limitations remain unresolved, and together they define the research gap addressed in this paper. First, the domain shift between simulated and measured data is severe: simulation data are a mathematical representation of physical processes, whereas measured data are affected by environmental noise, structural resonance and sensor errors, so the feature distributions of the two domains are almost completely separated. Recent transfer-learning studies have narrowed this gap for bearings and compressors [29–31], but they assume sufficient source data and a fault pattern that remains stable across domains, which does not hold for the weak and gradually evolving impacts of a loosening connecting nut that are buried in strong low-frequency excitation. Second, variable speed and pressure conditions require a generalization capability beyond the training scope; existing domain generalization strategies such as MixStyle [32] mix feature statistics at a single level and were not designed for the scale separation between incipient and severe loosening. Third, the cross-entropy classifier ignores the ordinal nature of the fault levels; although ordinal regression has been applied to general ordinal classification tasks [33], it has not been used to encode the four-level physical progression of nut loosening. These three limitations are addressed in this paper respectively by a validated multibody source combined with CORAL-based second-order statistic alignment, a shallow-stage MixStyle module, and a weighted ordinal head, which are integrated into the Swin-MCO model.

Table 1 compares representative studies in terms of method, data source, working conditions and reported performance.

Table 1. Comparison of representative studies on cross-domain fault diagnosis of threaded connections and rotating machinery.

Study	Method	Data and Object	Working Conditions	Reported Result	Main Limitation
Chakrapani et al. [28]	Transfer of AlexNet, GoogLeNet, ResNet-50 and VGG networks with fine-tuning	Vibration signals of a single-stage single-acting air compressor	One machine, five fault types	100% with AlexNet (570 s)	Needs labeled data of the same machine; no domain alignment; no severity grading
Tian and Xu [19]	Spatio-temporal deep belief network (STF-DBN)	Multi-source signals; TE benchmark and an offshore compressor	Single machine, several fault types	Effective detection and early warning; no single accuracy reported	Same-domain only; no cross-domain alignment
Wang et al. [17]	SE-ResNet-50 feature extraction with Swin Transformer classification	Vibration of a cement mill roller press and a main exhaust fan	Field data, normal and fault states, 6:2:2 split	99.21% (roller press) and 98.89% (exhaust fan)	Binary normal/fault only; single plant; no cross-domain alignment
Bo et al. [16]	Residual vision Transformer (BoltResViT)	Images of railway bolts	Field images acquired by a robot	99.28% monitoring accuracy	Vision based; binary looseness; no severity grading
Shi et al. [29]	Wasserstein distance based multi-scale adversarial adaptation	C-MAPSS aero-engine degradation data	Cross-domain RUL prediction	RMSE 11.84 and 12.01, score 16.62 and 18.21	Predicts remaining life rather than severity level
Zhou et al. [32]	MixStyle statistic mixing	PACS, VLCS and Office-Home	Unseen domains (leave-one-domain-out)	82.3% on PACS vs. 79.5% for ResNet-18	General vision benchmark; single-level statistics; not fault specific
This work	Swin-MCO: multi-scale Swin, MixStyle, Deep CORAL and weighted ordinal head	Simulated and bench data of a crosshead connecting nut	25 speed and pressure conditions, four severity levels	95.80/95.57/96.72%	Validated on one test rig; preload estimated from torque

To address the above issues, this paper proposes the Swin-MCO model. Although Deep CORAL, MixStyle and ordinal regression have each been used in fault diagnosis, this work does not treat them as a simple stack. For the specific failure mechanism of crosshead connecting nut loosening, three coupled design choices are made. First, a multi-scale fusion head is adopted to match the physical scale separation between the weak, high-order impacts of incipient loosening and the low-order, high-energy impacts of severe loosening. Second, MixStyle is inserted only after the shallow stages, where the feature statistics mainly encode style shifts caused by speed and pressure, while the spatial content still retains the fault impacts. Third, a weighted ordinal head is used to encode the physical progression of the fault from normal through slight and moderate to severe. These three choices operate at different levels, namely the feature scale, the shallow style statistics and the output label structure, and they are therefore complementary rather than redundant, as quantified by the module-wise ablations and baseline comparisons in Section 5. In addition, the source domain is not an unvalidated simulator output: the rigid-flexible coupled multibody model is checked against bench measurements, and only 20% of the experimental data are labeled for training to mimic the scarcity of labeled data in industrial environments. To our knowledge, this is the first framework that combines a multibody simulation source, multi-scale Swin features, second-order statistic alignment and ordinal regression for cross-domain connecting nut loosening diagnosis.

2. Data Acquisition and Signal Preprocessing

2.1. Dynamic Simulation

To obtain fault data under multiple operating conditions, this paper establishes a rigid-flexible coupled multibody dynamic simulation model of a reciprocating compressor. The piston rod, crosshead and connecting nut are modeled as flexible bodies, and the mesh in the thread area is locally refined to 1 mm. A nonlinear spring-damper model based on the Hertz contact theory is adopted as the contact force model. Six pairs of collision contacts are set up, including connecting rod–crosshead, crosshead–connecting nut and crosshead–slideway. Fixed joints, revolute joints and parallel joints are configured according to the actual connection relationships. The model is shown in Figure 1.

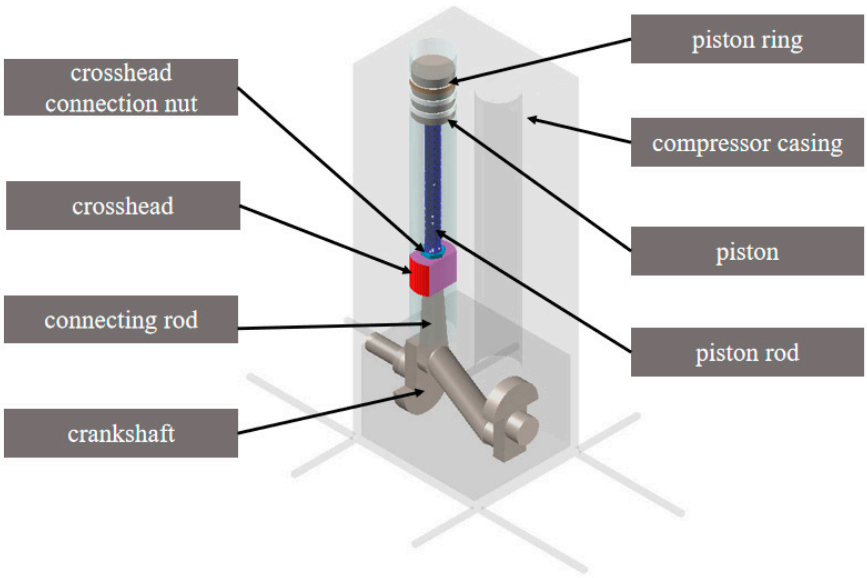


Figure 1. Multibody dynamics software model.

The material properties, kinematic joints and contact parameters of the model are listed in Tables 2–4, respectively. The piston rod, crosshead and connecting nut were modeled as flexible bodies with tetrahedral elements (maximum size 10 mm, minimum size 4 mm), and the thread regions were locally refined to 1 mm; the chordal deviation was set to 0.1 and small features were removed to avoid distorted elements. Contact face sets (Patch Sets) and multi-point constraint (FDR) elements were used to transfer loads between the flexible bodies and the rigid components.

Table 2. Material properties of the main components of the multibody model.

Component	Material	Elastic Modulus E (N/mm^2)	Density (kg/mm^3)	Poisson's Ratio
Crankshaft	45 steel	2.06×10^5	7.85×10^{-6}	0.30
Connecting rod	45 steel	2.06×10^5	7.85×10^{-6}	0.30
Crosshead	QT600-3	1.69×10^5	7.10×10^{-6}	0.26
Connecting nut	40Cr	2.11×10^5	7.82×10^{-6}	0.28
Piston rod	35CrMo	2.10×10^5	7.83×10^{-6}	0.28
Piston assembly	45 steel	2.06×10^5	7.85×10^{-6}	0.30
Piston ring	Filled PTFE	0.88×10^5	2.20×10^{-6}	0.40
Cylinder block	QT450-10	1.73×10^5	7.00×10^{-6}	0.27

Table 3. Kinematic joint definitions of the multibody model.

Component I	Component II	Joint Type
Cylinder block	Foundation	Fixed joint
Cylinder block	Crankshaft	Revolute joint
Crankshaft	Connecting rod	Revolute joint
Connecting rod	Crosshead	Parallel joint + contact
Crosshead	Piston rod	Contact
Crosshead	Connecting nut	Contact
Crosshead	Slideway	Contact
Connecting nut	Piston rod	Contact
Piston rod	Piston	Fixed joint
Piston ring	Cylinder wall	Contact

Table 4. Contact parameters of the main contact pairs.

Contact Pair	Stiffness K (N/mm^n)	Damping C ($N{\cdot}s/mm$)	Force Exponent n	Boundary Penetration dmax (mm)	Dynamic Friction Coefficient μd
Connecting rod and crosshead	5.0×10^5	3000	1.5	0.01	0.05
Crosshead and piston rod	2.1×10^6	2000	2.2	0.005	0.1
Crosshead and connecting nut	2.1×10^6	2000	2.2	0.005	0.1
Crosshead and slideway	1.0×10^5	1000	1.5	0.01	0.08
Connecting nut and piston rod	2.1×10^6	2000	2.2	0.005	0.1
Piston ring and cylinder wall	5.0×10^5	500	1.2	0.02	0.05

The loads and boundary conditions were set to reproduce the actual machine. The cylinder pressure measured on the test bench was imported and applied to the piston end face by Akima spline interpolation, so that the gas excitation follows the real pressure profile rather than an idealized one. To reproduce the belt-driven crankshaft, a rotational drive function was applied to both ends of the crankshaft, for example 20π rad/s for 600 rpm. The simulation time was determined by the rotational speed, and the time step was set according to the sampling frequency of 25,600 Hz, so that the simulated signals have the same temporal resolution as the measured ones.

The equations of motion were integrated in Adams, and the numerical settings were chosen so that the contact forces at the impact instants are resolved. The integration step was fixed at 3.906×10^{-5} s, i.e., one tenth of the sampling interval of 25,600 Hz, and the solver used the GSTIFF integrator with an error tolerance of 1×10^{-4}. The integration step is one tenth of the acquisition interval, so that the contact events are resolved at the sampling resolution of the measurement. The maximum penetration of each contact pair remained below the maximum damping depth listed in Table 4, so that the Hertzian contact law was not over-driven.

The initial pre-tightening torque of the connecting nut was 40 N·m, which was sequentially reduced to 38 N·m (slight loosening), 34 N·m (moderate loosening) and 28 N·m (severe loosening). The step function was used to apply and then remove the pre-tightening torque to simulate the loosening state. The simulation rotational speeds were 480 rpm, 540 rpm and 600 rpm, and the exhaust pressures were 0.1 MPa, 0.2 MPa and 0.3 MPa. In total, 9 standard working condition combinations were considered. A total of 600 cycles were collected for each working condition and each loosening state, resulting in 21,600 simulation samples in total. For each condition and state, 700 cycles were computed, and the first 100 cycles, which contain the start-up transient caused by preload application, were discarded. The simulation results showed that the acceleration amplitude increased systematically as the pre-tightening torque decreased: the amplitude was within 10,000 mm/s^2 under the normal state, approximately 15,000 mm/s^2 under slight loosening, exceeded 20,000 mm/s^2 under moderate loosening, and rose above 40,000 mm/s^2 under severe loosening. The vibration signals of the crosshead connecting nut are shown in Figure 2.

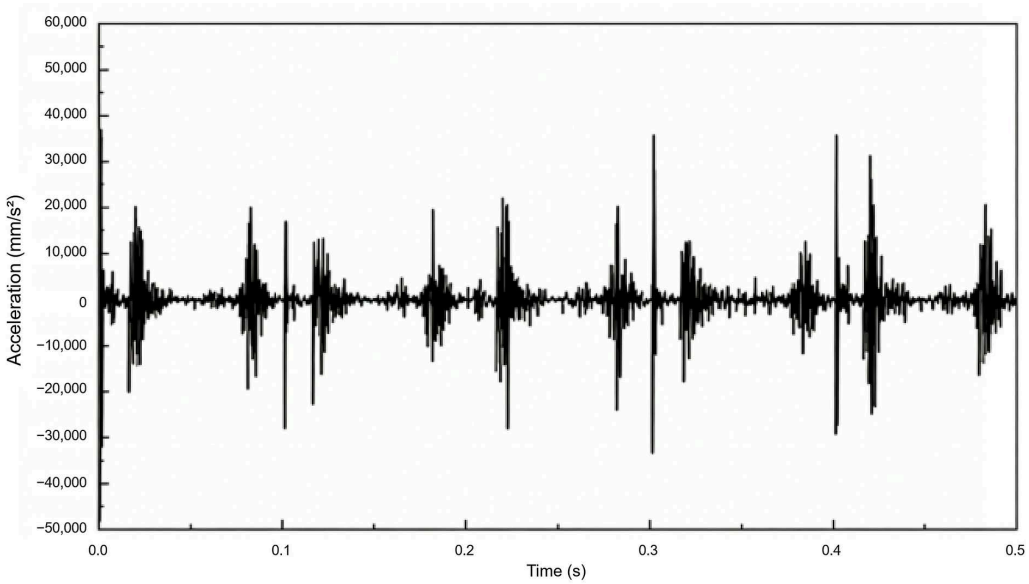


Figure 2. Vibration signal of crosshead connecting nut.

The four levels were defined by the torque setpoint rather than by a directly measured preload. A fully released nut was re-tightened to 40, 38, 34 and 28 N·m with a digital torque wrench, corresponding to 100%, 95%, 85% and 70% of the nominal torque. These setpoints correspond, through T = KFd, to approximately the same percentages of the initial clamping force, and they are placed around the two reference levels reported for bolted joints [34]: a reduction of the clamping force to about 80% of its initial value marks the entry into the structural loosening stage, and a reduction to about 70% has been used as the criterion of full loosening. Because the tightening torque and the preload are related by T = KFd, with K ≈ 0.2 for steel threads and d = 20 mm (M20 × 2.5 thread: internal diameter 20 mm, pitch 2.5 mm), the nominal preloads are approximately 10.0, 9.5, 8.5 and 7.0 kN. These four values are well below the yield limit of the nut, so the joint remains in the elastic range at every level. They are nevertheless estimates, since the scatter of the friction coefficient can change the preload at the same torque, and the torque is therefore reported as the controlled experimental variable throughout this paper. Each state was reproduced by releasing and re-tightening the nut in the same sequence to ensure repeatability. The four levels and the corresponding estimated preloads are summarized in Table 5. The effective preload was not measured independently, which is stated as a limitation in Section 7.

Table 5. Torque levels and estimated preloads. The preloads are estimated from T = KFd with K ≈ 0.2 and d = 20 mm; they were not measured directly, and the torque is the controlled experimental variable.

State	Torque Setpoint (N·m)	Ratio to Nominal Torque	Estimated Preload (kN)	Loosening Basis
Normal	40	100%	10.0	Initial pre-tightening state
Slight loosening	38	95%	9.5	Early stage of clamping-force decrease
Moderate loosening	34	85%	8.5	≈80% of initial clamping force: structural loosening stage [34]
Severe loosening	28	70%	7.0	≈70% of initial clamping force: full-loosening criterion [34]

The four states are defined by the torque applied at assembly rather than by the residual torque measured after loosening. First, the nominal torque of the connecting nut is 40 N·m, the value specified for this machine, and it defines the normal state. Second, the three loosened states are set at 95%, 85% and 70% of the nominal torque, so that the two lower setpoints bracket the 80% and 70% clamping-force reference levels reported for bolted joints [34] and the four levels follow an established degradation sequence rather than an arbitrary spacing. Third, the torque is applied at every assembly instead of being measured after the event, so the level of each sample is known exactly; a residual-torque definition would require measuring the torque that remains in a joint which has already rotated, and that measurement cannot be made without disturbing the state that is to be diagnosed. For every state the nut was fully released and the threads were cleaned before the nut was re-tightened with a calibrated digital torque wrench, in the same sequence and to the same setpoint, so that all specimens share an identical assembly history; the same sequence of tightening, loosening and cleaning was repeated for every state and every condition. The effective preload was not measured by an independent method. The values in Table 5 are obtained from T = KFd with the standard coefficient K ≈ 0.2 for as-received steel threads and are therefore nominal estimates, because the friction coefficient varies with lubrication, surface finish and re-use. For this reason the torque, and not the estimated preload, is treated as the controlled variable. The physical ordering of the four states is nevertheless supported by the measurements: the peak acceleration, the RMS value and the energy in

the 6–9 kHz band all increase monotonically as the torque decreases (Section 2.4), which is the response expected from a progressively loosened joint. An ultrasonic time-of-flight measurement or a load washer beneath the nut would provide an independent check of the preload, and this is listed as a direction for future work in Section 7.

2.2. Experimental Platform and Datasets

A vertical two-cylinder reciprocating compressor fault simulation test bench was built, which consists of a variable frequency motor, crankshaft, connecting rod, crosshead, connecting nut, piston assembly, cylinder, air storage tank and other components. The vibration acceleration sensor was mounted at the center point of the outer wall of the first-stage crosshead–slideway, the key-phase sensor was installed on the flywheel protective cover, and the high-frequency dynamic pressure sensor was arranged at the center of the air valve of the first-stage cylinder. Data acquisition was implemented via the NI-USB6218 data acquisition card, with the sampling rate of vibration acceleration set to 25,600 Hz.

A high-precision digital torque wrench (Figure 3) was used to quantitatively reproduce four loosening states: normal at 40 N·m, slight loosening at 38 N·m, moderate loosening at 34 N·m, and severe loosening at 28 N·m. For each adjustment the compressor was stopped and powered off, the side panel above the piston rod was removed, the nut surface was cleaned and the locking washer was released so that the nut could be turned; after each setting a short trial run was performed to confirm normal operation, and the data were recorded only after this check. To verify the cross-domain generalization performance of the diagnostic model, on the basis of 9 standard operating conditions with rotating speeds ranging from 480 rpm to 600 rpm and exhaust pressure from 0.1 MPa to 0.3 MPa, three types of non-standard operating conditions were added: 6 variable-speed operating conditions at 510 rpm and 570 rpm, 6 variable-pressure operating conditions at 0.15 MPa and 0.25 MPa, and 4 dual-variable operating conditions with both rotating speed and pressure changed. In total, there were 25 operating conditions. For each condition, 400 steady-state cycles were collected, yielding 40,000 experimental samples in total. The samples were divided into four datasets by operating condition type: Dataset A for standard operating conditions with 14,400 samples, Dataset B for variable-speed operating conditions with 9600 samples, Dataset C for variable-pressure operating conditions with 9600 samples, and Dataset D for dual-variable operating conditions with 6400 samples.

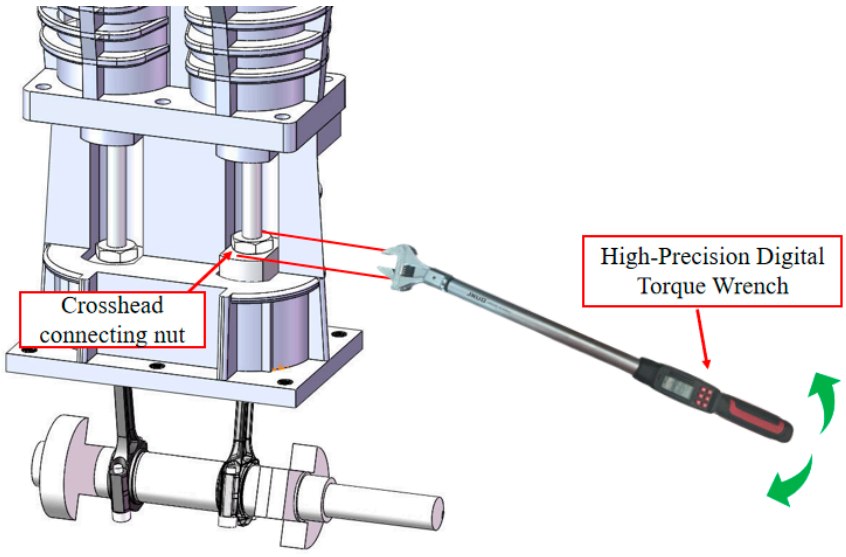


Figure 3. Position of the crosshead connecting nut and the high-precision digital torque wrench used to set its pre-tightening torque to the four levels of 40, 38, 34 and 28 N·m.

Through the preprocessing chain described in Section 2.3 and summarized in Table 6, the original vibration signals were converted into 224 × 224 two-dimensional texture images, which served as the standardized input of the diagnostic model.

Table 6. Parameters of the signal preprocessing chain.

Parameter	Value	Remark
Sampling rate	25,600 Hz	Consistent with the simulation step setting
Angular points per revolution	M = 3600	0.1° resolution; maximum 3200 points per cycle at 480 rpm
SSA embedding window	L = 50	50 reconstructed components
Component screening weights	Kurtosis 0.7	Kurtosis 0.7, energy 0.2 and correlation 0.1, consistent with the implementation
Adaptive threshold	$T = \mu s + \lambda \sigma s$	0.524 for the illustrated cycle
Image size	224 × 224	PAA-Max with 224 segments, about 1.6° of crank angle per pixel
SSA effectiveness	kurtosis from 6.49 to 10.36	Computation time about 0.07 s per cycle

2.3. Signal Preprocessing and Image Encoding

The complete chain, from the angular-domain sequence to the encoded image, is shown in Figure 4. Each cycle was triggered by the key-phase pulse and resampled in the angular domain. A linear time-to-angle mapping followed by cubic-spline interpolation produced M = 3600 points per revolution, i.e., an angular resolution of 0.1°, which exceeds the maximum of 3200 points acquired per cycle at 480 rpm and 25,600 Hz.

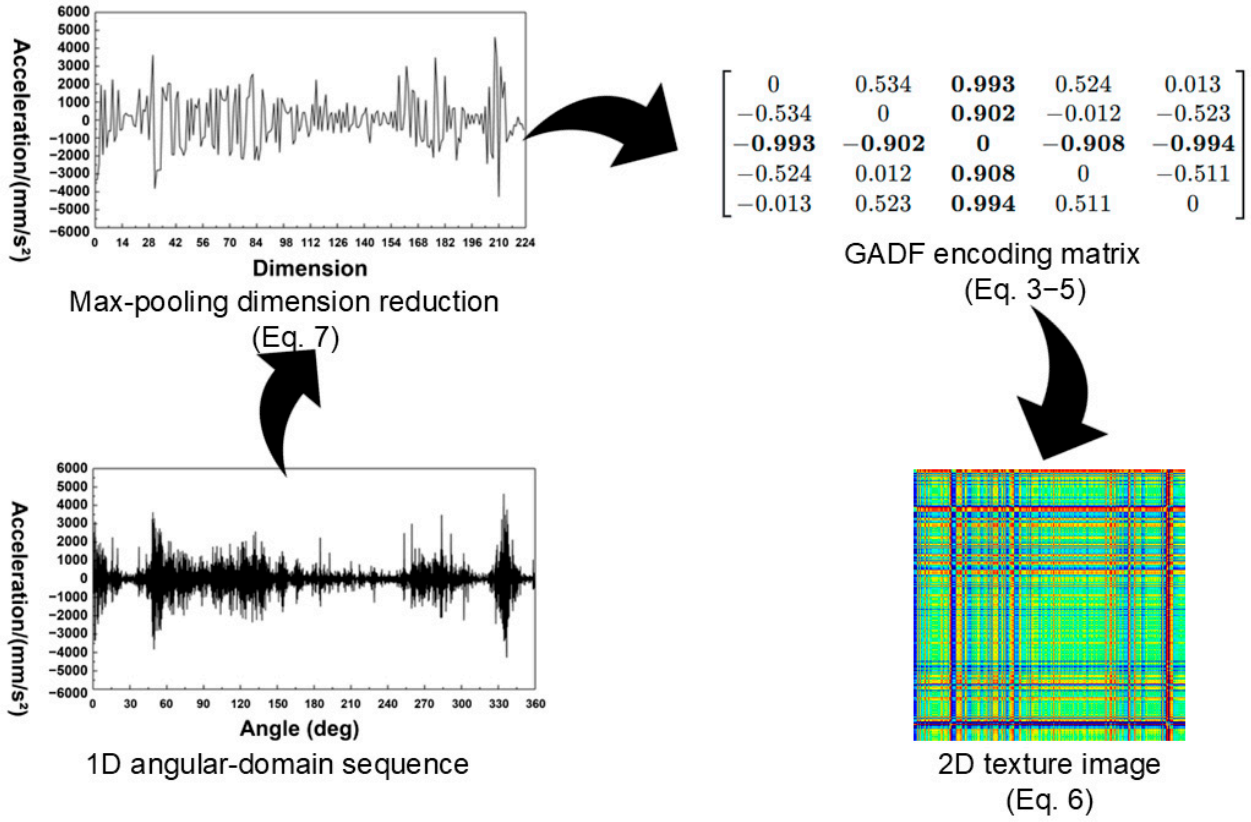


Figure 4. Gramian angular difference field process based on PAA-Max.

The resampled sequence was decomposed by improved singular spectrum analysis (SSA) with an embedding window L = 50, giving 50 reconstructed components. Each component i was scored by Si = 0.7Ki + 0.2Ei + 0.1ρi, where Ki, Ei and ρi are the kurtosis, the normalized RMS energy and the Pearson correlation coefficient with the original sequence, respectively; the three indices were normalized by the range method, and the kurtosis weight was set to 0.7, and the energy and correlation weights were set to 0.2 and 0.1 by comparing the diagnostic accuracy over multiple conditions. Components whose scores were above the adaptive threshold T = μs + λσs were retained and reconstructed, where μs and σs are the mean value and the standard deviation of the scores of the 50 components and λ = 1 is a coefficient that controls the strictness of the screening; for the illustrated cycle, T = 0.524. With this screening, the kurtosis of the reconstructed signal increases from 6.49 to 10.36 and the computation time is about 0.07 s per cycle.

$$S_i = 0.7K_i + 0.2E_i + 0.1\rho_i,\ i = 1,2,\ldots,L \tag{1}$$

$$T = \mu_s + \lambda\sigma_s \tag{2}$$

The reconstructed sequence x = {x1, ..., xN} was first normalized to the range $[-1, 1]$ by min-max normalization, because the polar encoding used in the next step requires the argument of the arccos function to lie within $[-1, 1]$. Each normalized sample was then mapped to the polar angle φi = arccos(xi) and the radius ri = ti/N, where ti is the time index, so that the temporal order is retained in the radial coordinate, and the Gramian angular difference field matrix was obtained from the trigonometric difference of the angles of every pair of samples. Because the sine kernel responds to the rate of change of the angle, the encoding is sensitive to the transient impacts that characterize the loosening fault.

$$\tilde{x}_i = 2(x_i - x_{min})/(x_{max} - x_{min}) - 1 \tag{3}$$

$$\varphi_i = \arccos(\tilde{x}_i), -1 \le \tilde{x}_i \le 1 \tag{4}$$

$$r_i = t_i/N \tag{5}$$

$$G = \sin(\varphi_i - \varphi_j),\ i, j = 1, 2, \ldots, M \tag{6}$$

To match the input of the Swin Transformer, piecewise aggregation with maximum pooling (PAA-Max) divided the 3600-point sequence into 224 equal segments and retained the absolute peak of each segment, so that each pixel aggregates about 16 sampling points, corresponding to approximately 1.6° of crank angle. This preserves the transient impacts that would be smoothed by mean pooling. The size of 224 was chosen because it matches the 4 × 4 patch embedding of the Swin Transformer, is compatible with ImageNet pretrained weights, and gave the best accuracy and computation trade-off among the candidate input sizes of 128, 224 and 320. Compared with mean aggregation, the maximum-pooling strategy raises the Gini coefficient of the encoded image from 0.087 to 0.190, indicating stronger preservation of transient impacts.

$$x_p = max_{j\in\Omega_p}|x_j|, p = 1, 2, \ldots, 224 \tag{7}$$

2.4. Validation of the Multibody Model

The multibody model was validated against bench measurements at 600 rpm and 0.2 MPa, a mid-range condition representative of normal operation. In the time domain, the simulated peak acceleration increases systematically as the torque decreases: below 10,000 mm/s^2 at 40 N·m, about 15,000 mm/s^2 at 38 N·m, above 20,000 mm/s^2 at 34 N·m when the crosshead strikes the slideway, and above 40,000 mm/s^2 at 28 N·m; the measured signals show the same macroscopic evolution. The impact epochs also coincide: the shocks

at 0.01 s and 0.055 s correspond to the reversal of the combined piston force, and the shocks at 0.02 s and 0.08 s correspond to the crosshead striking the slideway. The measured signals additionally exhibit a transient at about 0.07 s under 34 N·m and 28 N·m, which is caused by the multi-source excitation of the test bench.

In the frequency domain, with a sampling rate of 25,600 Hz the effective analysis band is 0 to 12,800 Hz. The simulated spectral peak amplitude rises from about 200 mm/s^2 at 40 N·m to 600, 1200 and 2000 mm/s^2 as the torque decreases, and the dominant energy band migrates from 2 to 3 kHz to 4 to 6 kHz and 6 to 9 kHz; the measured spectra show the same migration, with stronger broadband energy in the 2 to 3 kHz band due to the multi-source impacts of the test bench. The agreement in amplitude evolution and spectral migration indicates that the multibody model reproduces the load-excitation mechanism of the loosening fault, so the simulated signals are suitable as the source domain. Table 7 summarizes the comparison, with the detailed data presented in the Appendix A Figures A1 and A2.

Table 7. Qualitative validation of the multibody model against bench measurements (600 rpm, 0.2 MPa).

State	Torque (N·m)	Simulated Peak (mm/s^2)	Simulated Spectral Peak (mm/s^2)	Dominant Band
Normal	40	below 10,000	approximately 200	uniform spectrum
Slight	38	approximately 15,000	approximately 600	2 to 3 kHz
Moderate	34	above 20,000	approximately 1200	4 to 6 kHz
Severe	28	above 40,000	approximately 2000	6 to 9 kHz

The residual difference is concentrated in the high-frequency band, where the simplified treatment of the valve and the estimated preloads of Table 5 are the main sources of error; both are stated as limitations in Section 7. Because the dominant band and its migration are reproduced, the multibody model is used as the source domain of the transfer task, and a quantitative spectral error metric is left to future work.

3. Deep-Transfer-Learning Methodology

3.1. Deep CORAL

In the development of deep network adaptation, DDC (Deep Domain Confusion), proposed by Tzeng et al. from the University of California, Berkeley [35], is the first deep network adaptation algorithm. It adopts the AlexNet network as the feature extraction model and applies the standard MMD metric criterion to realize multi-domain adaptation. The CORAL method, put forward by Sun et al. [36] in 2016, is an unsupervised domain adaptation approach improved on the basis of DDC. This method uses linear transformation to align the covariance distributions of second-order statistical features in both the source and target domains, and it is a shallow non-end-to-end training method. In the same year, to introduce the idea of CORAL into deep neural networks, Sun et al. proposed Deep CORAL, a deep nonlinear end-to-end method. Instead of searching for a linear transformation matrix, it converts covariance alignment into a loss function. By embedding the loss of Deep CORAL into the deep learning network, the network can learn domain-invariant nonlinear features through back propagation, thus achieving end-to-end domain adaptation. The basic principles of Deep CORAL are as follows:

Suppose the source domain is $D_S = \{x_i^S\}_{i=1}^{n_s}$. The target domain is $D_T = \{x_j^T\}_{j=1}^{n_T}$. The source domain feature matrix of a certain layer in the network is $H_S \in \mathbb{R}^{n_S \times d}$. The target domain feature matrix is $H_T \in \mathbb{R}^{n_T \times d}$. The feature dimension is d.

First, calculate the sum of the covariance matrices of the source domain C_S and the target domain C_T, as shown in Formulas (8) and (9).

$$C_S = \frac{1}{n_S - 1}\left(H_S^T H_S - \frac{1}{n_S}\left(\mathbf{1}^T H_S\right)^T\left(\mathbf{1}^T H_S\right)\right) \tag{8}$$

$$C_T = \frac{1}{n_T - 1}\left(H_T^T H_T - \frac{1}{n_T}\left(\mathbf{1}^T H_T\right)^T\left(\mathbf{1}^T H_T\right)\right) \tag{9}$$

where **1** is a column vector in which all values are 1. Equations (8) and (9) compute the covariance matrices of the source and target domains, respectively, with the same expression.

Deep CORAL defines a distance loss function based on the Frobenius norm to measure the difference in feature distributions between the source domain and the target domain, as shown in Equation (10).

$$L_{\text{CORAL}} = \frac{1}{4d^2}\|C_S - C_T\|_F^2 = \frac{1}{4d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(C_S^{ij} - C_T^{ij}\right)^2 \tag{10}$$

where $\|\cdot\|_F$ is the Frobenius norm of the matrix. In Equations (8)–(10), H_S and H_T are the source and target feature matrices, each with n_S or n_T rows and d columns; C_S and C_T are d by d symmetric covariance matrices; the superscript T denotes transposition; and the loss in Equation (10) follows Sun et al. [36].

3.2. MixStyle

Domain generalization (DG) technology learns from source domains to construct networks capable of adapting to any unseen domains, which are then embedded into deep learning models to improve the generalization performance of deep learning networks when dealing with unseen domains. Proposed by Zhou et al. in 2021 [32], MixStyle is a domain generalization method based on the random mixing of feature statistics. Its core idea draws on style transfer theory: the content information of an image is reflected in the spatial structure of feature maps, while the style information is determined by the global statistics of feature maps. The fundamental principles of MixStyle are presented as follows:

First, set the dimensions of the feature map for the input data $B \times C \times H \times W$, then calculate the mean $\mu(x)_{i,c}$ and standard deviation $\sigma(x)_{i,c}$.

$$\mu(x)_{i,c} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{i,c,h,w} \tag{11}$$

$$\sigma(x)_{i,c} = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_{i,c,h,w} - \mu(x)_{i,c}\right)^2 + \epsilon} \tag{12}$$

where ε is a small positive constant used for numerical stability, B denotes the batch size, C is the number of channels, H is the height of the feature map, W is the width of the feature map, i indexes the sample within the mini-batch, c indexes the channel, and $x_{i,c,h,w}$ is the value of the feature map of sample i at channel c and spatial position (h, w). Accordingly, $\mu(x)_{i,c}$ and $\sigma(x)_{i,c}$ are scalars computed separately for each channel of each sample over the H by W grid, so that Equations (11) and (12) give the channel-wise mean and standard deviation of each feature map. The constant ε keeps the argument of the square root strictly positive, so that the standard deviation never vanishes and the normalization in Equation (15) remains well-defined.

To simulate data under unknown working conditions, the MixStyle domain generalization strategy shuffles the data within each batch. It adopts random mixing coefficients from the Beta distribution $\lambda \in \text{Beta}(\alpha,\alpha)$ to perform linear interpolation on the original and

shuffled data, so as to obtain mixed style statistics. The calculation formulas are shown in Equations (13) and (14).

$$\gamma_{\text{mix}} = \lambda\sigma(x) + (1-\lambda)\sigma(\tilde{x}) \tag{13}$$

$$\beta_{\text{mix}} = \lambda\mu(x) + (1-\lambda)\mu(\tilde{x}) \tag{14}$$

where λ is the mixing style proportional coefficient, α is the hyperparameter of the Beta distribution, and α is set to 0.1 in this paper. The mixing operation described by Equations (13)–(15) follows Zhou et al. [32]. In both equations, $\tilde{x}i = x\pi(i)$ denotes the feature map of the sample that a random permutation π of the current mini-batch assigns to position i, so the second term is the statistic of a different real sample and each mixed statistic combines the style statistics of two different samples. The permutation is redrawn for every mini-batch and every forward pass, so the paired sample, and hence the mixed style, varies over training.

Next, the mixed feature statistics are used to normalize and reconstruct the original feature map x. The original style is first removed while the content is retained, and the mixed style is then recombined with the content, so that the output carries statistics interpolated between two samples while its spatial content is preserved. The calculation formula is shown in Equation (15). Equation (15) does not create a new physical signal: it produces a feature-level interpolation between two real samples of the source domain.

To make the mixing operation explicit: for each mini-batch, a random permutation π of the batch samples is drawn, and sample i takes the shuffled feature map of sample $\pi(i)$. A mixing coefficient λ is drawn independently for each sample from the Beta distribution $\text{Beta}(\alpha, \alpha)$, and the same λ is used for both the mixed mean and the mixed standard deviation in Equations (13) and (14). Equation (15) then replaces the style of each sample while its content is preserved. The module is embedded after Stage 1 and Stage 2, where the statistics mainly encode style shifts caused by rotating speed and pressure while the spatial content retains the fault impacts. It is applied with a probability of 0.5 during training and disabled during inference. MixStyle therefore does not synthesize new physical signals; it interpolates feature statistics and acts as a regularizer that improves the robustness to unseen speed and pressure combinations, as confirmed by the ablation study in Section 5.1. Because the two statistics in Equations (13) and (14) are taken from real samples of the source domain, the mixed style always lies between the styles of two observed samples, and the module therefore widens the style coverage of the source domain instead of producing physically impossible operating conditions. The hyperparameter $\alpha = 0.1$ makes the Beta distribution strongly bimodal, so that λ is close to 0 or 1 and the mixed statistics stay close to one of the two samples rather than collapsing to their average, which preserves the fault-related content. The activation probability $p = 0.5$ leaves half of the forward passes unperturbed, which stabilizes training and keeps the ordinal structure of the features.

$$\text{MixStyle}(x) = \gamma_{\text{mix},i,c} \cdot \frac{x - \mu(x)}{\sigma(x)} + \beta_{\text{mix},i,c} \tag{15}$$

3.3. Ordinal Regression Loss Function

The loosening fault of the connecting nut on the crosshead of a reciprocating compressor follows a physical evolution pattern, progressing from normal status to slight, moderate and finally severe loosening. To capture this physical variation, the ordinal regression loss converts the multi-class classification task into multiple binary subtasks, where the fault category is determined by judging whether the fault degree exceeds a specific level without reaching a more severe one.

Assume the fault category is $y \in \{0,1,\ldots,K-1\}$, $K = 4$. They are classified into normal, slight, moderate and severe. Thus, the task is split into three binary classification subtasks.

Subtask 1: Determine whether the fault level of the data is higher than normal. If yes, proceed to Subtask 2: Determine whether the fault level of the data is higher than slight. If yes, proceed to Subtask 3: Determine whether the fault level of the data is higher than moderate. Then encode the determined fault levels into binary labels with clear hierarchical relationships. For example, the binary label for moderate fault data is {1,1,0}. Next, construct the weighted ordinal regression loss function, as shown in Formula (16).

$$L_{OR} = -\sum_{k=0}^{K-2} w_k [t_k log(\hat{p}_k) + (1 - t_k) log(1 - \hat{p}_k)] \quad (16)$$

where $\hat{p}_k$ is the predicted probability of the k-th subtask of the network; ω_k is the class weight coefficient. Here k indexes the binary subtask, K is the number of severity levels and equals four, so that the summation over k from 0 to K − 2 covers the three binary subtasks, and the formulation follows Cheng et al. [33].

Finally, sum the predicted values of all binary classification subtasks for the data to obtain the final fault category. The calculation formula is shown in Equation (17). Equation (16) gives the weighted loss of a single binary subtask and Equation (17) sums the predictions of all subtasks.

$$\hat{y} = \sum_{k=0}^{K-2} \mathbb{I}(\hat{p}_k > 0.5) \quad (17)$$

where $\mathbb{I}(\cdot)$ is an indicator function, which takes the value 1 if the condition in the parentheses is satisfied, otherwise it takes the value 0. Accordingly, the value computed by Equation (17) is an integer class index ranging from 0 to K − 1, where 0 denotes the normal state and K − 1 denotes the severe state, so that the four levels are ordered as normal, slight, moderate and severe.

4. Swin-MCO Model Architecture

4.1. Overall Structure

In view of the above limitations, this paper proposes a deep ordered transfer diagnostic model named Swin-MCO Taking the Swin Transformer as the backbone network for feature extraction, the method integrates domain adaptation, domain generalization and ordinal classification, realized by the MixStyle, Deep CORAL and ordinal regression loss modules, to address the coupled challenges of cross-domain discrepancy, variable-condition generalization and ordered fault severity learning. It thus realizes the accurate diagnosis of connecting nut loosening faults across different domains and variable working conditions. The architecture of the proposed fault diagnosis model is shown in Figure 5.

Relative to the standard Swin-T backbone, three modifications are made, and no other change is applied to the backbone. First, the single-scale classification head is removed and replaced by a two-scale fusion head that combines the Stage 3 and Stage 4 feature maps (Section 4.1), so that the weak high-order impacts of incipient loosening and the low-order high-energy impacts of severe loosening are represented at their own resolution. Second, MixStyle layers are inserted after Stage 1 and Stage 2 (Section 3.2), where the feature statistics mainly encode the style induced by speed and pressure. Third, the Softmax layer and the cross-entropy loss are replaced by an ordinal head with K − 1 = 3 binary outputs trained with the weighted loss of Equation (16) (Section 3.3). The patch embedding, the shifted-window attention blocks and the stage depths follow the original Swin-T configuration, so the contribution of this work lies in the three modifications and in their interaction rather than in the backbone itself.

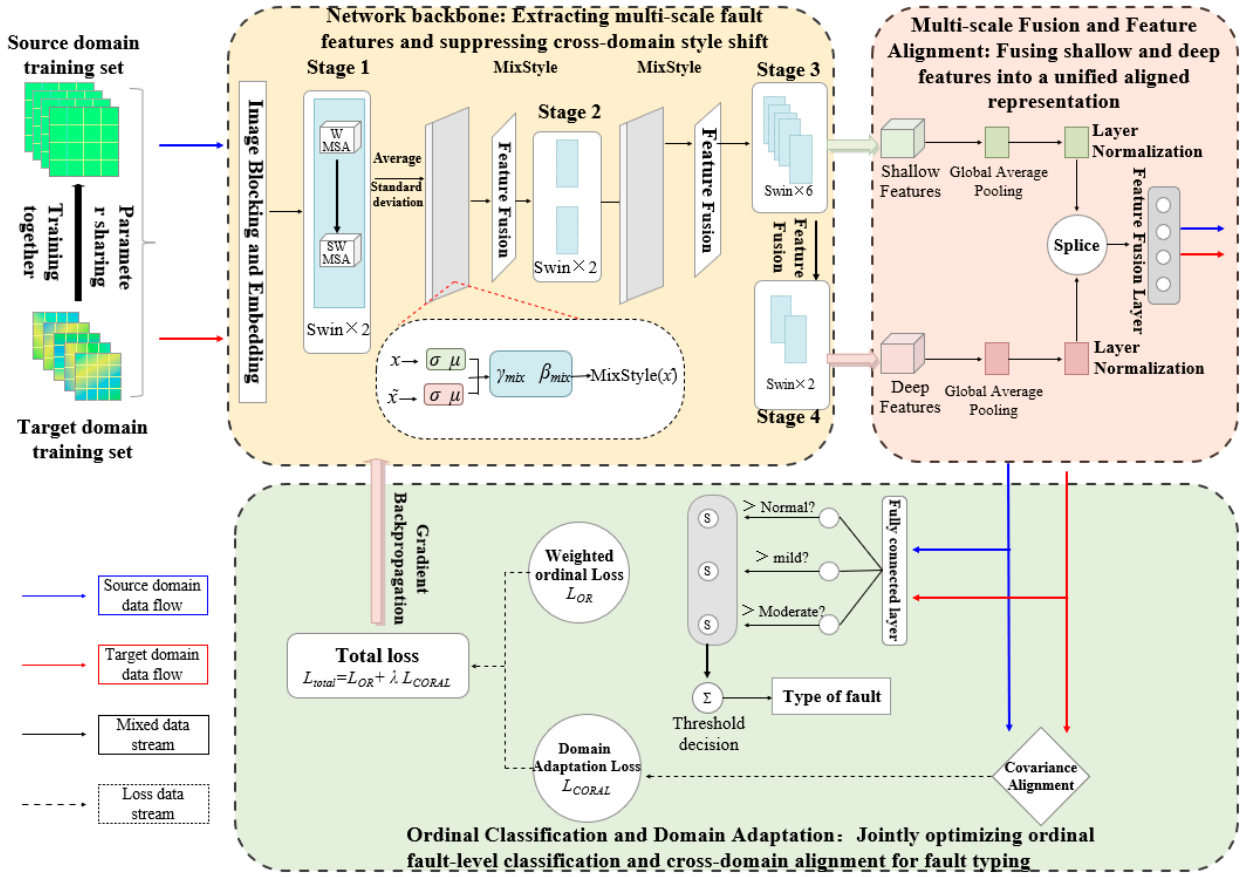


Figure 5. Swin-MCO model.

(1) Main Body of the Model Network

The MixStyle domain generalization module is introduced into the shallow feature extraction stage of the Swin Transformer network and embedded after Stage 1 and Stage 2. At this stage, 2D texture images are decoupled into content information and style information. The content information mainly contains fault impact signals generated by loosening connecting nuts, while the style information reflects differences in the statistical distribution of data caused by changes in rotating speed and pressure. During training, the MixStyle module calculates the mean and standard deviation of the input data flow. It adopts random coefficients to linearly mix the style statistics of different samples to generate virtual style features that simulate unknown working conditions, which are then re-fused with the content information. The augmented data adapted to various working conditions are fed into the next stage of the network. This module focuses on learning key fault features that are independent of working condition changes. After multiple rounds of forward and backward propagation training, it can effectively perform fault diagnosis under variable working conditions.

(2) Multi-scale Fusion and Feature Alignment

Fault features present multi-scale characteristics. Features of minor faults manifest as high-order transient impacts, while those of severe faults appear as prominent low-order high-energy impacts. The conventional Swin Transformer adopts sequential single-scale feature extraction across stages, which struggles to adapt to variations in the order and pattern of signals. Accordingly, this paper proposes a multi-scale fusion module. At Stage 3 of the network, the number of feature channels is 384 with limited downsampling operations. The feature maps of mixed data streams maintain a high spatial resolution, and a smaller receptive field is adopted to extract microscopic fault textures from the feature maps and capture weak early fault signals. At Stage 4 of the network, the number of feature channels reaches 768.

After processing, the compressed feature maps carry global information, enabling analysis of the macroscopic energy distribution that indicates fault deterioration. The shallow features extracted from Stage 3 and deep features from Stage 4 are converted into compressed vectors via global average pooling and then normalized, before being concatenated into a hybrid vector with 1152 dimensions. Ultimately, this vector is mapped to a 768-dimensional vector through the feature fusion layer.

The rationale of this design is as follows. Physically, the impacts of incipient loosening are weak, high-order and short-lived, whereas severe loosening produces low-order, high-energy impacts; the former requires high spatial resolution and a small receptive field, while the latter requires a large receptive field and global energy statistics. Structurally, the Stage 3 feature maps (384 channels) preserve local texture details, while the Stage 4 feature maps (768 channels) encode global information; pooling and concatenating the two levels yields a 1152-dimensional vector, which is mapped to 768 dimensions so that the subsequent CORAL alignment and ordinal head operate on a unified representation. Because the fusion is performed before the CORAL loss, the alignment is constrained by both scales and cannot be dominated by the high-energy low-order impacts of severe states, which prevents the weak incipient features from being aligned away. Removing this module and using only the Stage 4 vector lowers the mean diagnosis accuracy, which confirms that the gain comes from preserving the high-resolution texture rather than from merely adding parameters.

(3) Ordinal Classification and Domain Adaptation

The mixed data stream after feature fusion is fed into the ordinal classification and domain adaptation module. This module calculates the covariance matrices of the source domain data stream and the target domain data stream respectively, and computes the Frobenius norm distance between them to obtain the Deep CORAL loss. It also calculates the fault type loss for the two data streams, and predicts fault severity intervals in three ordinal levels. The output of each node is processed by the Sigmoid activation function before the weighted ordinal loss is computed. The total loss of the model is the weighted sum of the domain adaptation loss and the weighted ordinal loss, as shown in Equation (18). The main network parameters are updated in an end-to-end manner via backpropagation of gradients to minimize the total loss and train a network model with high diagnostic accuracy.

$$L_{\text{total}} = L_{\text{OR}} + \lambda L_{\text{CORAL}} \tag{18}$$

where λ is the loss weight coefficient. λ L_{ord} is the weighted ordinal loss given in Equation (17), L_{CORAL} is the domain adaptation loss given in Equation (10), and λ balances the two terms; its value is 2.0, as listed in Table 8.

Table 8. Training hyperparameters of the Swin-MCO model.

Parameter	Value
Initial learning rate	0.00005
Learning rate schedule	cosine annealing (minimum 1×10^{-6})
Optimizer	AdamW
Weight decay	0.05
Batch size	128
Loss weight coefficient	2.0
Training epochs	40

4.2. Parameter Settings

The main structural parameters of the Swin-MCO model are determined based on the data characteristics of the fault signals generated by loosening connecting nuts on the crosshead of the reciprocating compressor, as shown in Table 9. The initial learning rate of the model is 0.00005, and the cosine annealing scheduler is adopted with a minimum

learning rate of 10^{-6}. AdamW is used as the optimizer, with a weight decay coefficient of 0.05, a batch size of 128, a loss weight coefficient of 2.0, and a total of 40 training epochs.

Table 9. Main structural parameters of the model.

Network Module	Layer Name	Core Parameters & Operations	Activation Function
Feature Extraction Network	Image Partition & Embedding	Kernel 4 × 4, Stride 4	-
	Stage 1	Swin Block ×2, Heads = 3	GELU
	MixStyle1	$p = 0.5$, $\alpha = 0.1$	-
	Feature Fusion Layer 1	Kernel 2 × 2, Stride 2	-
	Stage 2	Swin Block ×2, Heads = 6	GELU
	MixStyle2	$p = 0.5$, $\alpha = 0.1$	-
	Feature Fusion Layer 2	Kernel 2 × 2, Stride 2	-
	Stage 3	Swin Block ×6, Heads = 12	GELU
	Feature Fusion Layer 3	Kernel 2 × 2, Stride 2	-
	Stage 4	Swin Block ×2, Heads = 24	GELU
Domain Adaptation Module	Covariance Alignment Layer	Frobenius Norm Distance	-
Multi-Scale Fusion Module	Global Pooling 1	GAP, Stage 3	-
	Global Pooling 2	GAP, Stage 4	-
	Feature Concatenation Layer	384 + 768	-
	Feature Fusion Layer 4	Dropout, $p = 0.2$	GELU
Ordinal Classification Module	Ordinal Fully Connected Layer	768 → 3	-
	Binary Classification Output Layer	3 Binary Subtasks	Sigmoid
	Weighted Ordinal Loss	Binary Encoding, Weighted Loss	-

4.3. Transfer Task Configuration

The dataset partition for the transfer diagnosis task is specified as follows. All 21,600 simulation samples serve as the source-domain training set. Experimental Dataset A contains 14,400 samples (9 standard conditions × 4 states × 400 cycles), i.e., 3600 samples per state, and a class-stratified split was applied: 20% of Dataset A (720 samples per state, 2880 in total) formed the target-domain training set, and the remaining 80% (2880 samples per state, 11,520 in total) formed the validation set, as summarized in Table 10. The 2:8 ratio therefore refers to the split of Dataset A, not to the ratio between the source domain and the target domain. Dataset B (variable rotating speed), Dataset C (variable pressure) and Dataset D (both variable rotating speed and pressure) were used only as test sets and were never involved in training or model selection. No sample therefore appears in more than one subset, and all reported test accuracies are obtained on conditions unseen during training. This partition scheme simulates the actual industrial scenario in which labeled fault data are scarce.

The three test sets are designed so that generalization to unseen conditions can be tested on the bench, and the sense in which a condition is unseen must be stated precisely. Here a condition is unseen when its speed–pressure pair was not used in training: set B introduces 510 and 570 rpm, which lie between the training speeds of 480, 540 and 600 rpm; set C introduces 0.15 and 0.25 MPa, which lie between the training pressures of 0.1, 0.2 and 0.3 MPa; and set D introduces both at once Table 10 . The three sets therefore probe interpolation inside the operating envelope of the machine rather than extrapolation outside it, and interpolation is the regime that a field machine actually visits, because the speed follows the load and the discharge pressure follows the demand continuously. The tests

are not trivial for that reason. The source domain is simulated and the target domain is measured, and the Frobenius distance between their feature covariances before training is 47.28 (Section 5.1), so the model does not face near-identical data; and the three test sets are drawn from three different condition families, yet one trained model reaches 95.57 to 96.72% on all of them, with every error confined to adjacent severity levels. What these results establish is generalization across operating conditions on one machine; generalization across machine structure is a different question that the bench cannot answer and that is discussed in Section 5.3.

Table 10. Composition of the simulation and experimental datasets (four loosening states per condition).

Dataset	Conditions	Speed (rpm)	Pressure (MPa)	Cycles Per State	Samples	Role
Simulation (source)	9	480/540/600	0.1/0.2/0.3	600	21,600	training
Experimental A	9	480/540/600	0.1/0.2/0.3	400	14,400	20% training, 80% validation
Experimental B	6	510/570	0.1/0.2/0.3	400	9600	test
Experimental C	6	480/540/600	0.15/0.25	400	9600	test
Experimental D	4	510/570	0.15/0.25	400	6400	test
All experimental	25	—	—	—	40,000	—

Each sample is one complete steady-state cycle, and the cycles are non-overlapping in time, so no signal segment is shared between the training and validation sets. Because the cycles of each condition come from a single continuous recording, a session-level split based on independent re-installations was not applicable to the present dataset; this limitation is stated in Section 7.

The workflow of the transfer fault diagnosis method based on the Swin-MCO model is illustrated in Figure 6, which is generally divided into two phases: offline training and real-time diagnosis. In the offline training phase, first, the source domain training data under standard operating conditions are obtained via the rigid-flexible coupled multibody dynamics model. Meanwhile, target domain data under corresponding operating conditions are collected on the test bench. The collected data are converted into two-dimensional texture images through the preprocessing methods described in Section 2.3. Next, model parameters are initialized and pretrained weights from ImageNet are loaded. The image data sequentially pass through the feature extraction network, the MixStyle working condition enhancement module and the multi-scale fusion layer to generate high-dimensional feature vectors. The total loss is calculated by combining the ordinal regression loss and domain adaptation loss, and the network weights are updated using the backpropagation algorithm. The training process lasts for 40 epochs, and the model parameters with the best performance on the validation set are saved.

In the real-time diagnosis phase, on-site vibration data are collected in real time with the reciprocating compressor condition monitoring system, and converted into two-dimensional images after preprocessing to serve as the test input of the target domain. After loading the optimal model weights, the random mixing operation of the MixStyle module is disabled to fix the network parameters. After the model infers the target domain images, the ordinal regression classification head outputs the prediction probabilities of three binary classification subtasks. The final fault type is obtained through summary calculation, realizing the real-time diagnosis of the loosening fault of the crosshead connecting nut under variable operating conditions.

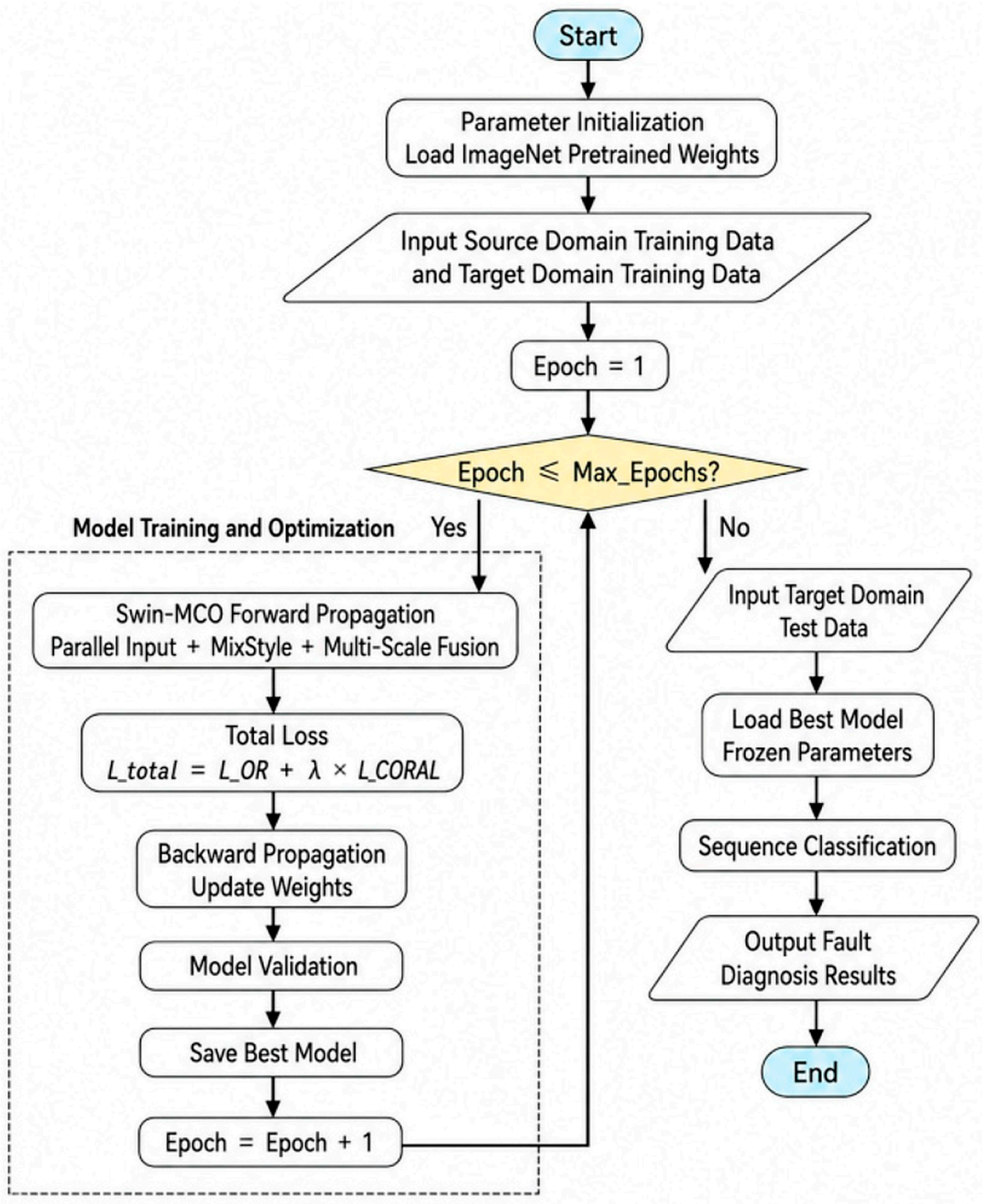


Figure 6. Fault diagnosis flow chart.

4.4. Complementarity of the Components

The two statistic-based modules do not compete, because they act on statistics of different scope and at different depths. MixStyle perturbs the first- and second-order statistics of individual feature maps in the shallow stages, and it acts only during training. Deep CORAL aligns the covariance of the final feature representation between the two

domains, and it acts in every forward pass, including inference. In the shallow layers the spatial content still carries the impulses produced by the loosening, whereas the statistics are dominated by speed and pressure, so modifying the statistics there changes the style without destroying the fault evidence. In the final layer the spatial arrangement has already been abstracted, so aligning the covariance there removes the residual domain shift. Used alone, CORAL must align two distributions in which the source side contains a single style per working condition, and the alignment is then dominated by the high-energy low-frequency impacts of the severe state. Used alone, MixStyle enlarges the style coverage of the source domain but enforces no relation between the source and the target distributions. Combined, MixStyle supplies the style diversity that CORAL requires, and CORAL supplies the explicit cross-domain constraint that MixStyle lacks.

Each element of the framework therefore removes a limitation that the others leave open. A plain Swin classifier is trained at a single scale with a nominal label space, so it cannot separate the weak high-order impacts of incipient loosening from the low-order impacts of the severe state, and it treats the four states as unrelated categories; the multi-scale fusion head and the ordinal head remove these two limitations, respectively. Deep CORAL removes the domain shift, but without MixStyle its alignment is driven by the dominant severe-state impacts, and without the fusion head the weak early-stage features would be aligned at the same scale as the high-energy features and would be suppressed. Ordinal regression preserves the order of the states, but it provides neither a mechanism against domain shift nor a defense against the imbalance between adjacent states; the CORAL term and the weighted binary loss cover these two gaps. The four elements therefore act on the feature scale, the shallow style statistics, the domain-level covariance and the output label structure, and none of them can replace another.

Table 1 places this work alongside representative studies. The Transformer-based diagnosis methods reported in [16,17] are developed and evaluated within a single domain, and the transfer-learning methods in [29–31] adopt convolutional backbones and treat the fault classes as nominal, unordered labels. A Transformer backbone has therefore not been combined with second-order statistic alignment and an ordinal output for reciprocating machinery, which is the combination examined here. Because those studies report results on datasets that differ in sensor placement, rotational speed and label definition, the quantitative comparison is made by re-implementing representative baselines on the same data (Table 11) rather than by transferring numbers across datasets.

Table 11. Diagnosis performance of the three models on test sets B, C and D.

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Kappa
ResNet18-DANN	B	52.65	62.30	52.65	54.47	0.369
ResNet18-DAN	B	67.44	75.19	67.44	68.76	0.566
Swin-MCO	B	95.80	95.82	95.80	95.80	0.944
ResNet18-DANN	C	52.87	60.65	52.87	54.55	0.372
ResNet18-DAN	C	66.37	72.73	66.37	67.50	0.552
Swin-MCO	C	95.57	95.60	95.57	95.57	0.941
ResNet18-DANN	D	50.35	59.77	50.35	52.08	0.338
ResNet18-DAN	D	67.40	75.31	67.40	68.85	0.565
Swin-MCO	D	96.72	96.76	96.72	96.72	0.956

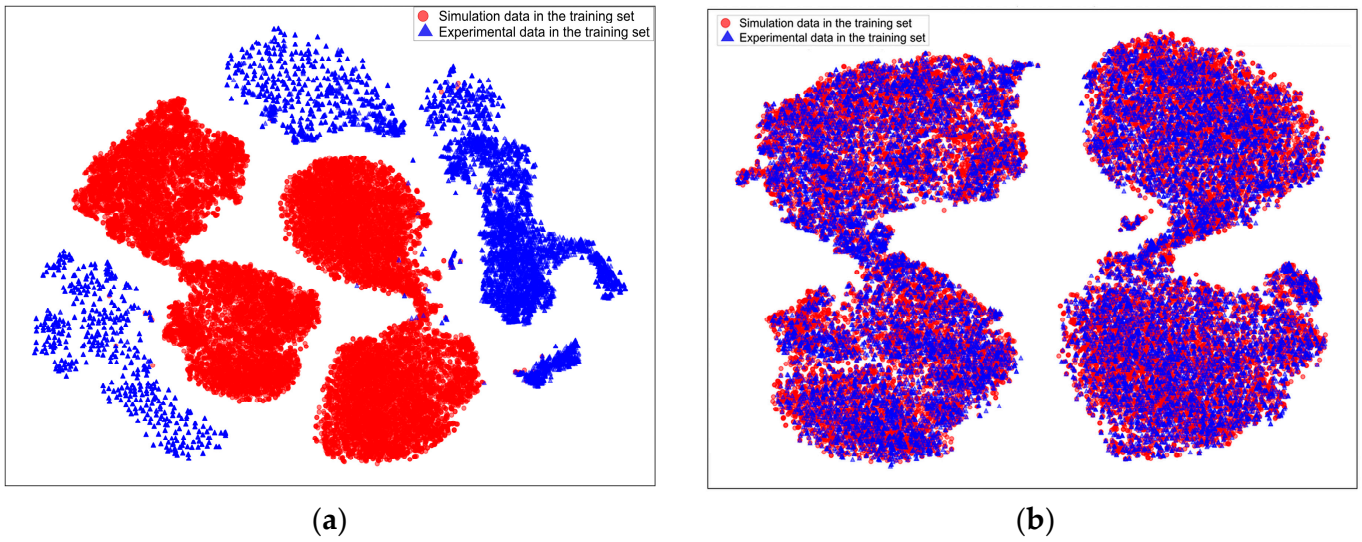
Accuracy, precision, recall and F1-score are expressed in percent (%); Kappa denotes Cohen's Kappa coefficient.

5. Experimental Verification

5.1. Ablation Study

To verify the effectiveness of each module in the Swin-MCO model, ablation experiments were conducted on the Deep CORAL domain adaptation module, the MixStyle domain generalization strategy and the ordinal regression loss function respectively.

(1) Effectiveness analysis of the Deep CORAL domain adaptation module. The t-SNE algorithm was adopted to reduce the features of the source domain and target domain training sets into a two-dimensional plane for visualization at three stages: before training, mid-training and late training, as shown in Figure 7.



(**a**) (**b**)

Figure 7. t-SNE visualization of feature distribution alignment between source domain and target domain. (**a**) Distribution difference of mid-training data between the source domain and the target domain. (**b**) Distribution difference of data in the later stage of training between the source domain and the target domain.

Meanwhile, the Frobenius norm distance of the covariance matrices between the two training sets was calculated as the quantitative evaluation index. Before training, the distribution distance between the source domain and the target domain was large, and the data of the target domain were excessively scattered, with the Frobenius norm distance reaching 47.2814. This indicated that there existed significant differences in statistical features between the two datasets. During mid-training, the source domain and the target domain gradually drew closer to each other, and the Frobenius norm distance dropped to 18.5633. It demonstrated that the model was learning domain discrepancies and adjusting the weight parameters of the feature extraction model via back propagation. In the late stage of training, four well-defined data clusters appeared in the figure, each corresponding to one type of looseness state: normal, slight looseness, moderate looseness and severe looseness. The two datasets achieved high overlap and alignment in the two-dimensional space, and the Frobenius norm distance decreased to 3.6013, a 92.4% reduction compared with the pre-training value. The results proved that the Deep CORAL loss effectively narrowed the feature distribution distance between the source domain and the target domain, and greatly improved the transferability of fault features.

(2) Effectiveness analysis of the MixStyle domain generalization strategy. In the Swin-MCO model, the activation probability p of the MixStyle module was set to 0.5. To verify the effectiveness of this module, a control model without the MixStyle module was constructed with $p = 0$, while all other network structures remained unchanged. Both models were trained for 40 epochs on the same training set and validation set, and their diagnostic

accuracy was subsequently tested using the target domain test set. The diagnostic results are presented in Figure 8.

Before the introduction of the domain generalization module, the model only learned fault features from the training set, achieving a maximum accuracy of 74.36% on the test set. After adopting the domain generalization module, the model obtained an accuracy of over 95% on the test set. This module enabled the model to extract fault features independent of working condition variations and capture the essential characteristics of faults. Meanwhile, it fused the mixed style statistics with content information to generate prediction data for various unknown working conditions, which greatly improved the model's accuracy in diagnosing faults under variable working conditions.

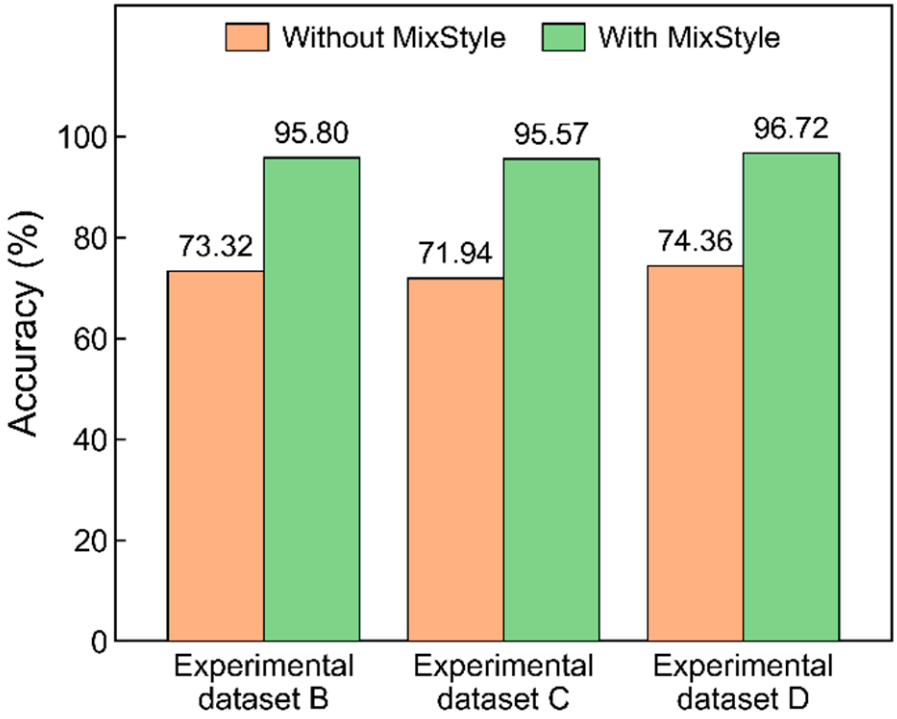


Figure 8. Results of domain generalization ablation experiments.

(3) Analysis of the effectiveness of the ordinal regression loss function. Figure 9 presents the fault diagnosis results of the Swin-MCO model on various datasets in the form of a confusion matrix. The diagnosis accuracy of the model on the target domain validation set was 95.00%, and the accuracy rates on the three target domain test sets were 95.80%, 95.57% and 96.72% respectively. The accuracy on the three types of test sets under unknown working conditions was comparable to that on the validation set, all staying above 95%. Notably, when deep learning models adopting the cross-entropy loss function were used to diagnose experimental Dataset D with simultaneous changes in rotating speed and pressure, a large number of normal states were misclassified as moderate faults. In contrast, misclassifications of the Swin-MCO model were strictly confined to adjacent fault types, and cross-level misjudgments were greatly mitigated. Analysis of the diagnosis results from other test datasets revealed that no cross-level misjudgment occurred with the proposed model in this paper. This indicated that the ordinal regression loss function effectively incorporated the physical evolution logic of fault severity into the model. The model captured the quantitative change relationship of gradual fault deterioration, and reduced the risks of severe missed alarms and false alarms during practical industrial applications.

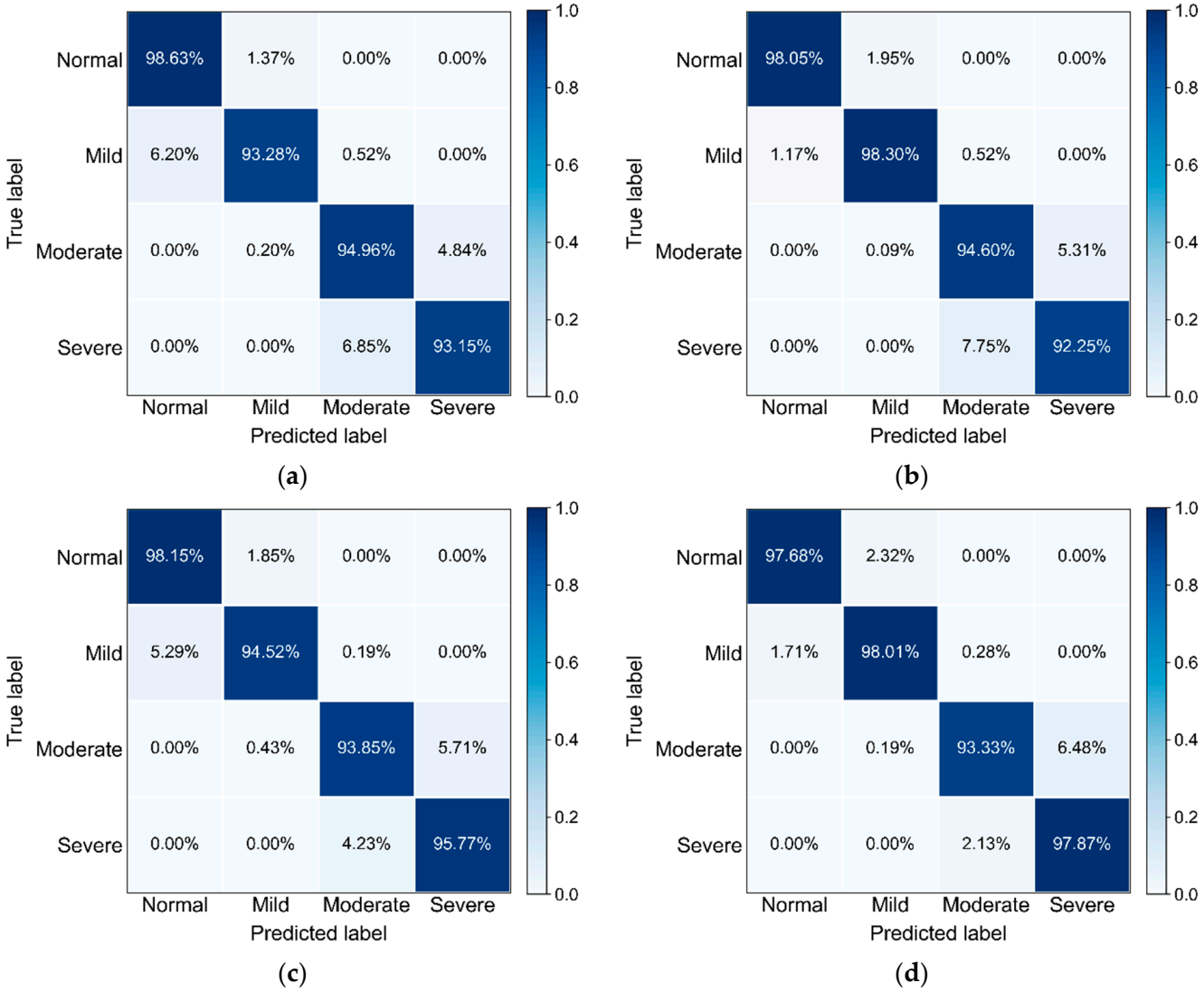


Figure 9. Confusion matrix for cross-working condition fault diagnosis. (**a**) Confusion matrix of validation set. (**b**) Confusion matrix of Experimental Dataset B. (**c**) Confusion matrix of Experimental Dataset C. (**d**) Confusion matrix of Experimental Dataset D.

5.2. Model Comparison and Analysis

To verify the superiority of the model proposed in this paper in fault diagnosis under variable working conditions and cross-domain scenarios, comparative analysis is conducted between the proposed model and two classic deep-transfer-learning algorithms, namely Domain-Adversarial Neural Network (DANN) and Deep Adaptation Network (DAN). DANN adopts mutual adversarial learning between the domain discriminator and the feature extractor, and realizes the alignment of feature distributions by misleading the discriminator. DAN introduces the Multi-Kernel Maximum Mean Discrepancy (MK-MMD) based on first-order statistics as the loss function. ResNet18 is used as the backbone network for feature extraction in both comparison methods. To ensure the fairness of comparative experiments, the three models share identical parameters: the initial learning rate is 0.00005, the batch size is 128, the optimizer is AdamW, and the number of training epochs is 40.

The accuracy, training loss and validation loss of the compared models are presented in Figures 10–12, respectively.

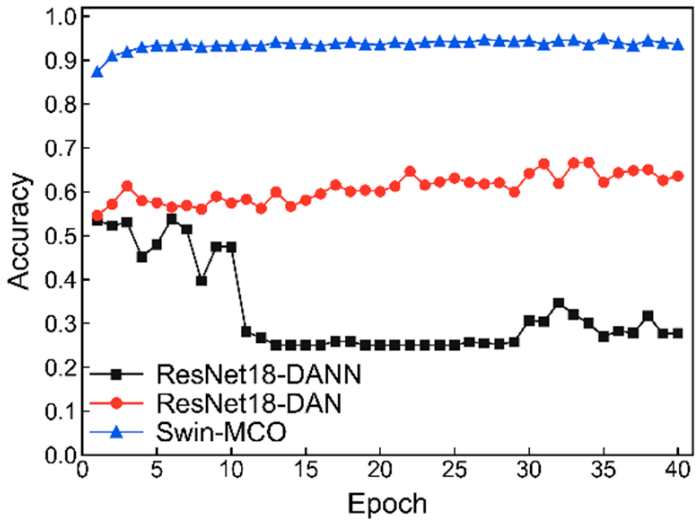


Figure 10. Accuracy comparison.

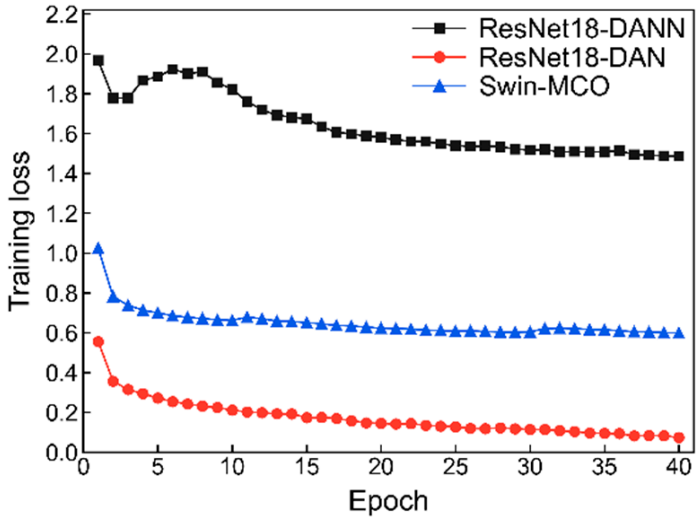


Figure 11. Training loss comparison.

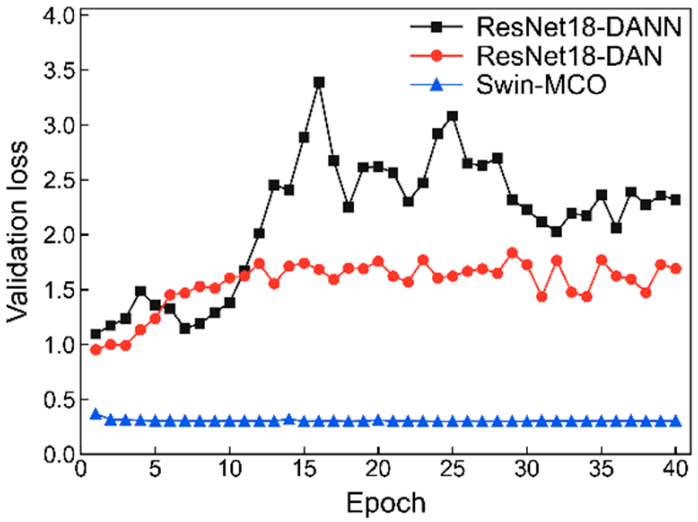


Figure 12. Validation loss comparison.

To simulate the application scenarios of online diagnosis for reciprocating compressors, five indicators including accuracy, precision, recall, F1-score and Kappa coefficient were adopted to evaluate the generalization ability of three models in diagnosing data under cross-domain and unknown operating conditions. A radar chart was also plotted for comparative analysis, as shown in Figure 13. The DANN model has the smallest covered area and the worst overall performance, while DAN outperforms DANN. The proposed method in this paper achieves over 90% for all evaluation indicators in the diagnosis of the loosening fault of the crosshead connecting nut of reciprocating compressors under variable operating conditions, demonstrating excellent diagnosis accuracy and cross-domain generalization capability.

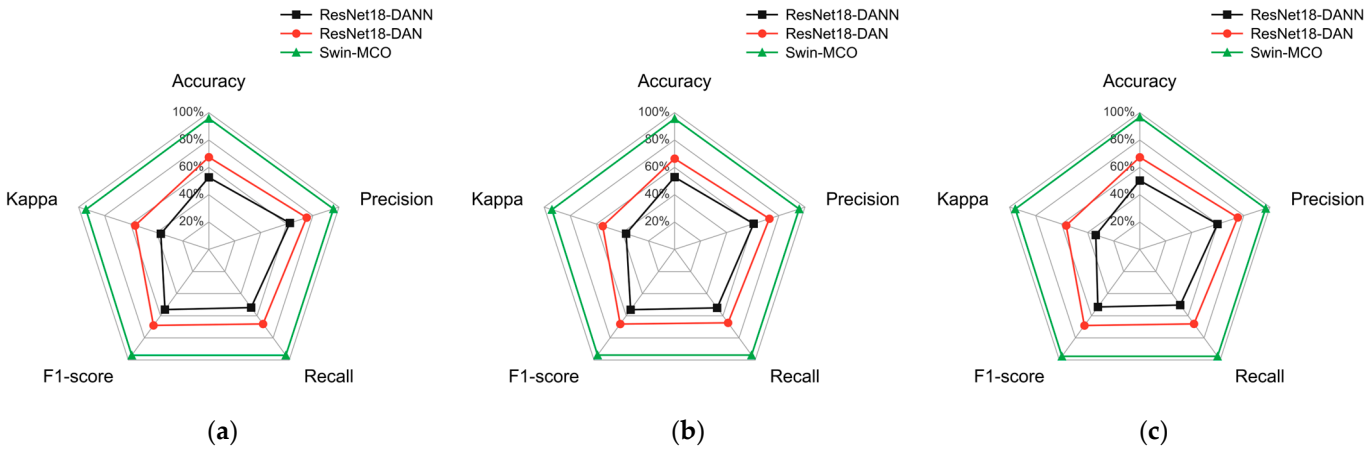


Figure 13. Radar chart of generalization performance for three models. (**a**) Experimental Dataset B. (**b**) Experimental Dataset C. (**c**) Experimental Dataset D.

The metric values plotted in the radar charts of Figure 13 are listed numerically in Table 11 to allow a direct comparison of the three models.

5.3. Computational Cost, Real-Time Feasibility and Industrial Transferability

The computational cost of the proposed model is analyzed here in the terms that matter for online monitoring: how much of one compression cycle the pipeline consumes and which stage dominates it. No benchmark of a specific industrial controller is claimed; the aim is to identify the cost structure of the chain and the components that limit a shorter update interval.

The backbone of Swin-MCO is Swin-T, which contains 28.3 M parameters and requires 4.5 GFLOPs for a 224 × 224 input [15]. The two-scale fusion head maps a 1152-dimensional fused vector back to 768 dimensions and the ordinal output layer consists of three binary units, so the parameter count of the complete model is dominated by the backbone. Stored in FP32, the weights amount to about 113 MB, which is modest for an industrial edge computer but far above the memory of a microcontroller-class device, and this fixes the class of hardware on which the model can be deployed. Both a single forward pass and the training of one mini-batch fit in the memory of a discrete GPU of the class used in this study.

The reference time scale of the diagnosis is one compression cycle. At 480 to 600 rpm one cycle lasts 0.10 to 0.125 s and contains 2560 to 3200 sampling points at 25.6 kHz, so signal acquisition is not the bottleneck of the chain. The dominant cost is the improved SSA screening, which requires about 0.07 s per cycle and therefore consumes 56 to 70% of one cycle; the angular resampling and the GADF encoding operate on the same cycle and are of the same order of magnitude. Because one cycle cannot therefore be processed within its

own duration on the computing platform used in this study, the diagnosis is issued on a buffered basis: the model is applied to the most recent cycle once the preceding one has been processed, so the update interval is a small integer number of cycles rather than one cycle. This is compatible with the application. Loosening develops over many operating hours, so a diagnosis delayed by a few cycles is still far ahead of the degradation, and the alarm does not have to be raised within the cycle in which the fault first appears. Two engineering routes shorten the interval further: replacing the SSA screening by band-pass filtering followed by envelope demodulation removes the dominant cost at some loss of impact fidelity, and deploying the network with FP16 or INT8 arithmetic reduces the network share well below the preprocessing share. The 224 × 224 input is fixed by the GADF encoding rather than by the Swin-T backbone, so a coarser angular grid would also reduce the network cost, at the price of a coarser texture.

The transferability of these results from the laboratory bench to an in-service industrial compressor deserves an explicit discussion. All experimental data in this study were acquired on a single laboratory bench, so the reported accuracies characterize that bench and not a production machine, and this is stated as the main limitation in Section 6. The simulation-to-bench transfer already spans one domain gap, because the source domain is a rigid-flexible coupled multibody model and the target domain is measured vibration, but the bench-to-field gap remains and cannot be closed by the present experiments. It has four components: the mounting position and the transmission path from the nut to the sensor differ, so the 4 to 9 kHz band that carries the loosening signature may be attenuated; adjacent machines, piping and valves add broadband background noise that is absent on the bench; the industrial machine differs in bore, stroke and frame stiffness, which shifts the impact spectrum; and the load and speed profiles vary continuously instead of being held at nine discrete setpoints. Three properties of the framework limit the resulting loss of performance. First, Deep CORAL aligns the second-order statistics of the two domains using unlabeled target data, so field vibration can be used for adaptation even though on-site fault labels are rarely archived. Second, MixStyle broadens the style coverage during training, so the network does not depend on the exact statistics of the bench. Third, the ordinal output confines errors to adjacent severity levels, so a shifted decision boundary degrades the severity estimate gradually instead of producing a wrong fault type. Closing the remaining gap requires a set of labeled field samples for fine-tuning, a recalibration of the two ordinal thresholds on field data, and the verification that the transmission path preserves the diagnostic band; this validation on in-service machines is planned and is stated in Section 7.

5.4. Discussion of the Misclassifications and of the Dual-Variable Results

The residual errors of Swin-MCO are not distributed uniformly over the four states, and their origin is physical rather than numerical. The four states differ only in the residual torque of the connecting nut, by 2, 4 and 6 N·m, and this difference reaches the sensor as a change in the amplitude and in the spectral position of the crosshead impact: the simulated spectral peak rises from about 200 to 600, 1200 and 2000 mm/s^2, while the dominant band migrates from 2 to 3 kHz toward 6 to 9 kHz. The first step, from 40 to 38 N·m, is the smallest in both amplitude and frequency shift, so the normal and slight images are the most similar pair in the dataset and are the pair that is confused most often, whereas the severe state is never assigned to the normal or the slight state. Three mechanisms account for the remaining adjacent-level errors. First, at the boundary between two levels the GADF texture changes only in the intensity of a few diagonals, which is a small perturbation of the 224 × 224 input, so a sample that lies close to the boundary can fall on either side. Second, the three binary subtasks of the ordinal head share one feature vector, so a marginal

violation of a single threshold produces an adjacent-level output by construction; this is the intended behavior of the ordinal formulation and converts what would be a distant error in a nominal classifier into a one-level shift. Third, the amplitude of the impact is modulated by the discharge pressure through the gas force, while its timing is modulated by the speed; angular resampling removes the timing modulation but not the amplitude modulation, so at the edges of the operating envelope an impact can appear one level stronger or weaker than it is. These effects are visible in the confusion matrices of Figure 9, in which the off-diagonal mass is concentrated between neighboring severity levels and no column is systematically over- or under-predicted.

Dataset D, in which speed and pressure change simultaneously, is the most demanding test of the framework, and it produced the highest accuracy of the three test sets, 96.72% against 95.80% on B and 95.57% on C. The advantage of Swin-MCO over the baselines is largest on this set: DANN falls to 50.35% and DAN to 67.40%, which is close to the level of a four-class random classifier, whereas the proposed model remains above 96%. This behavior is consistent with the mechanism of the two statistic-based modules. A joint change of speed and pressure alters the mean and the covariance of the shallow feature maps, that is, the style, and leaves the impact pattern that encodes the fault level largely intact once the angular resampling has removed the timing drift; MixStyle trains the network on styles interpolated between real samples, so styles lying between the observed ones are covered by construction, and Deep CORAL then aligns the remaining covariance mismatch between the two domains. The first-order alignment of DAN and the adversarial objective of DANN do not model this joint shift, which is why their accuracy collapses when the two variables move together. Two caveats are stated explicitly. The gap of about one percentage point between D and B or C is small, so the ranking of the three test sets is not interpreted as a property of the datasets. Dataset D also contains four conditions instead of six, so its within-class style diversity is smaller, which can raise the accuracy independently of the transfer difficulty. What these results support is therefore that the model degrades gracefully when both variables move, not that dual-variable conditions are intrinsically easier.

6. Conclusions

Based on the simulation and experimental data of crosshead connecting nut loosening under standard and unknown operating conditions, this paper constructed and verified the Swin-MCO deep ordered transfer diagnostic model. The main findings are as follows. (1) A rigid-flexible coupled multibody model was established, and the simulated signals agree with bench measurements in the evolution of acceleration amplitude and in impact epochs; the model generated 21,600 labeled samples, and 40,000 experimental samples were collected under 25 operating conditions. (2) The preprocessing chain of 3600-point angular resampling, improved singular spectrum analysis and PAA-Max GADF encoding converts each cycle into a 224 × 224 texture image while preserving the transient impacts. (3) Swin-MCO achieved accuracies of 95.80%, 95.57% and 96.72% on the variable-speed, variable-pressure and dual-variable test sets, and all misclassifications were confined to adjacent fault levels, so no cross-level error occurred. (4) Ablation studies confirmed that Deep CORAL, MixStyle and the ordinal regression loss each contribute to the mean accuracy, and the model outperforms DANN and DAN in accuracy, F1-score and Kappa coefficient. (5) The main limitation is that all experimental data were acquired on one laboratory bench with a single sensor layout, and the effective preload was estimated from torque rather than measured directly. (6) The computational cost of the chain was analyzed so that the real-time claim can be bounded: the improved SSA screening alone requires about 0.07 s, that is 56 to 70% of one compression cycle of 0.10 to 0.125 s, so the diagnosis is issued on a buffered basis over a small number of cycles, which is compatible with the

slow development of loosening; because all validation data were acquired on the laboratory bench, an on-machine validation remains necessary before field deployment.

7. Future Work

This work will be extended in four directions. First, the framework will be validated on in-service industrial reciprocating compressors, with long-term data covering different machine sizes, sensor locations and maintenance records, and the effective preload will be measured directly, for example by ultrasonic bolt-load measurement. Second, the diagnosis scope will be extended from nut loosening to coupled faults such as valve leakage, piston-ring wear and connecting-rod cracks, and to multi-label settings in which several faults coexist. Third, lightweight deployment will be pursued through structured pruning, knowledge distillation and quantization, and the diagnosis output will be linked to a maintenance decision rule so that the remaining useful life of the connection can be estimated. Fourth, physics-informed augmentation and uncertainty quantification will be introduced, so that the multibody model generates controlled fault severities while the model flags samples that fall outside the training distribution.

Author Contributions: Conceptualization, W.X.; Methodology, W.X.; Validation, Z.L.; Formal analysis, Z.L.; Data curation, Y.G.; Writing—original draft, W.X.; Writing—review and editing, Y.C.; Visualization, Y.G.; Resources, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Open Fund of the Key Laboratory of Advanced Nuclear Energy Design and Safety, Ministry of Education, grant number KLANETDS-KF-202510.

Data Availability Statement: The simulated and measured vibration signals, the condition labels and the data split of Table 10 are available from the corresponding author on reasonable request, because the raw records are large. To make the work reproducible, the multibody model parameters (Tables 2–4), the complete preprocessing chain (angular resampling, improved SSA and PAA-Max GADF encoding) and the training configuration (Table 9) are specified in the article. The PyTorch 2.8.0 implementation of Swin-MCO, together with the trained weights and the script that reproduces the splits, is available from the corresponding author on reasonable request.

Conflicts of Interest: Author Xidong Wang was employed by the PetroChina Tuha Oilfield Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The PetroChina Tuha Oilfield Company had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Figure A1 presents the simulated time-domain acceleration waveforms for four preload torques: at 40, 38, 34, and 28 N·m, the impact peaks are below 10,000 mm/s^2, approximately 15,000 mm/s^2, above 20,000 mm/s^2, and above 40,000 mm/s^2, respectively. A lower preload torque produces larger impact amplitudes and denser pulses. Figure A2 shows the corresponding simulated spectra: the spectral peaks for the four states are approximately 200, 600, 1200, and 2000 mm/s^2, respectively, with the energy always concentrated in the 0–3000 Hz low-frequency band.

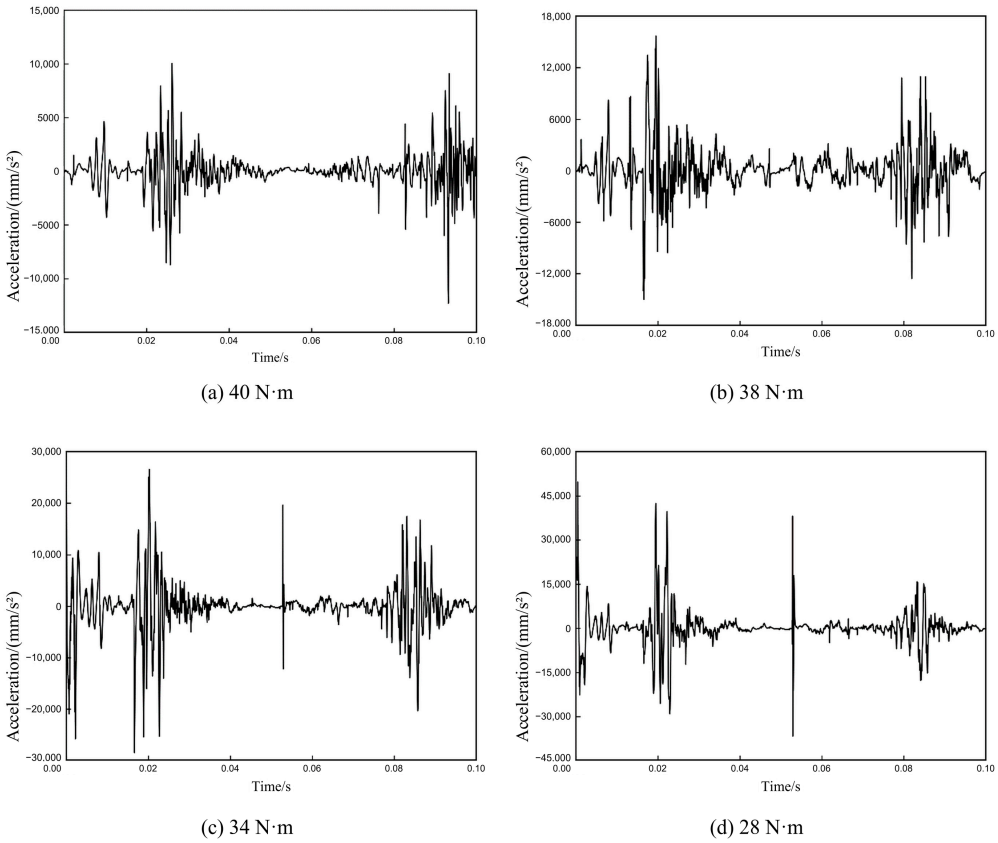


Figure A1. Time domain comparison diagram of simulation data.

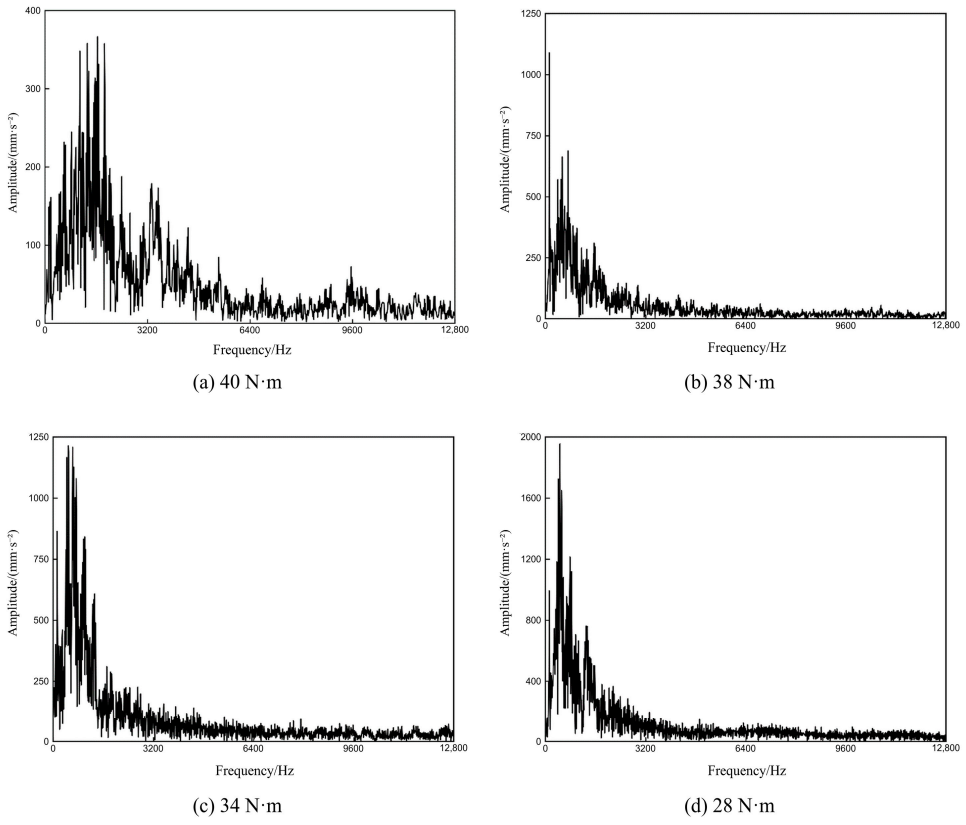


Figure A2. Frequency domain comparison diagram of simulation data.

References

1. Yu, L. Current Status and Development Trend of Fault Diagnosis Technology for Reciprocating Compressors. *Fluid Mach.* **2014**, *42*, 36–39.
2. Duan, Q.; Lu, R.; Xie, H.; Ping, J.; Lu, C.; Zhou, X.; Gao, J.; Li, J. Fault Diagnosis of Air Compressor in Nuclear Power Plant Based on Vibration Observation Window. *IEEE Access* **2020**, *8*, 222274–222284. [CrossRef]
3. Xiao, S.; Liu, S.; Cheng, S.; Xue, X.; Song, M.; Sun, X. Dynamic Analysis of Reciprocating Compressor with Clearance and Subsidence. *J. Vibroeng.* **2017**, *19*, 5061–5085. [CrossRef]
4. Wang, J.; Ma, Y.; Cheng, G.; Wang, Y.; Jiang, Z. Model Construction and Dynamic Analysis of Crankshaft of Reciprocating Compressor under Variable Working Conditions. *Fluid Mach.* **2022**, *50*, 28–34, 96. [CrossRef]
5. Lei, N.; Ding, H.; Tang, Y.; Miao, H.; Jiang, P.; Huang, F. Dynamic Response Analysis of Composite Clearance Faults in Transmission Mechanism of Reciprocating Compressor. *Noise Vib. Control* **2025**, *45*, 113–119.
6. Wei, L.; Tang, H.; Zhou, Y.; Wang, M. Fracture Failure Analysis of Piston Rod of Hydrogenation Circulating Compressor. *Hot Work. Technol.* **2023**, *52*, 157–160.
7. Li, X.; Diao, A.; Guo, Y.; Jia, X.; Zhang, C.; Peng, X. Quantitative Diagnosis of Loose Piston Rod Threads in Reciprocating Compressors for Hydrogen Storage and Transport. *Int. J. Hydrogen Energy* **2023**, *48*, 37013–37030. [CrossRef]
8. Gong, H.; Liu, J.; Feng, H. Research Review on Loosening Mechanism and Anti-Loosening Methods of Threaded Connections. *J. Mech. Eng.* **2022**, *58*, 326–347, 360. (In Chinese) [CrossRef]
9. You, L.; Liu, C.; Chen, H.; Fan, H. Research Progress on Influencing Factors of Bolted Joint Structure Loosening. *J. Phys. Conf. Ser.* **2024**, *2694*, 012062. [CrossRef]
10. Huang, H.; He, Q. Bolt Looseness Localization with Connection-Stiffness-Varying Flange. *Smart Mater. Struct.* **2024**, *33*, 095005. [CrossRef]
11. Du, J.; Qiu, Y.; Li, J. Research on Bolt Loosening Mechanism under Sine-on-Random Coupling Vibration Excitation. *Machines* **2025**, *13*, 80. [CrossRef]
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 770–778. [CrossRef]
14. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Hartwig, A. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
15. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 10012–10022. [CrossRef]
16. Bo, W.; Yuan, R.; Lv, Y.; Liu, H.; Hua, J.; Zhu, W. BoltResViT: An Enhanced Residual Vision Transformer for Robotic-Assisted Nondestructive Railway Bolt Looseness Monitoring. *Nondestruct. Test. Eval.* **2025**, *41*, 6862–6881. [CrossRef]
17. Wang, C.; Li, Y.; Wang, T.; Yang, C.; Liang, C. Bearing Fault Diagnosis Based on SE-ResNet-50 and Swin Transformer. *Modul. Mach. Tool Autom. Manuf. Tech.* **2026**, *8*, 177–182. (In Chinese)
18. Li, Y.; Hu, M.; Ma, B. Fault Diagnosis Method for Reciprocating Compressors Based on VSG-CNN. *Mod. Manuf. Eng.* **2025**, *541*, 138–147. (In Chinese)
19. Tian, H.; Xu, Q. A Spatio-Temporal Fault Diagnosis Method Based on STF-DBN for Reciprocating Compressor. *J. Intell. Manuf.* **2024**, *35*, 199–216. [CrossRef]
20. Zhao, H.; Huang, J.; Wang, J.; Wen, H.; Li, X. Improved SSA-VMD Algorithm and Its Application in Fault Diagnosis of Reciprocating Compressors. *Lubr. Eng.* **2022**, *47*, 147–152. (In Chinese)
21. Li, Y.; Wang, P.; Wu, S.; Ba, P. Fault Feature Extraction Method for Reciprocating Compressors Based on AO-VMD. *J. Mech. Electr. Eng.* **2023**, *40*, 673–681. (In Chinese)
22. Song, M.; Wang, J.; Zhao, H.; Li, Y. Reciprocating Compressor Valve Fault Diagnosis Based on Improved CEEMDAN and CHDE. *Mach. Tool Hydraul.* **2023**, *51*, 202–207. (In Chinese)
23. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35. [CrossRef]
24. Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Conditional Adversarial Domain Adaptation. In *Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018*; IEEE: New York, NY, USA, 2018; Volume 31, pp. 1640–1650. [CrossRef]
25. Long, M.; Cao, Y.; Wang, J.; Jordan, M.I. Learning Transferable Features with Deep Adaptation Networks. In *Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015*; ICML: San Diego, CA, USA, 2015; pp. 97–105. [CrossRef]
26. Sun, B.; Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 443–450.

27. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable Are Features in Deep Neural Networks? In *Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 27, pp. 3320–3328. [CrossRef]
28. Chakrapani, G.; Venkatesh, S.N.; Aravinth, S.; Sugumaran, V. Transfer Learning-Based Fault Diagnosis of Single-Stage Single-Acting Air Compressor. *J. Vib. Eng. Technol.* **2024**, *12*, 4411–4428. [CrossRef]
29. Shi, H.; Huang, C.; Zhang, X.; Zhao, J.; Li, S. Wasserstein Distance Based Multi-Scale Adversarial Domain Adaptation Method for Remaining Useful Life Prediction. *Appl. Intell.* **2023**, *53*, 3622–3637. [CrossRef]
30. Feng, T.; Qiu, L.; Wang, J.; Yang, B.; Zhang, Q. A Deep Domain-Adaptive Anomaly Detection Method for Scroll Compressor Speed Transfer Using Domain-Adversarial Training and Attention Mechanisms. *Results Eng.* **2025**, *28*, 108270. [CrossRef]
31. Ding, G.; Xu, Z.; Wang, Z. Cross-Condition Bearing Fault Diagnosis Method Based on Deep Transfer Metric Learning. *Modul. Mach. Tool Autom. Manuf. Tech.* **2025**, 177–183, 187. (In Chinese)
32. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. Domain Generalization with MixStyle. *arXiv* **2021**, arXiv:2104.02008. [CrossRef]
33. Cheng, J.; Wang, Z.; Pollastri, G. A Neural Network Approach to Ordinal Regression. In *Proceedings of the IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–6 June 2008*; IEEE: New York, NY, USA, 2008; pp. 1279–1284. [CrossRef]
34. Yang, G.; Yang, L.; Chen, J.; Xiao, S.; Jiang, S. Competitive Failure of Bolt Loosening and Fatigue under Different Preloads. *Chin. J. Mech. Eng.* **2021**, *34*, 141. [CrossRef]
35. Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; Darrell, T. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv* **2014**, arXiv:1412.3474. [CrossRef]
36. Sun, B.; Feng, J.; Saenko, K. Return of Frustratingly Easy Domain Adaptation. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016*; AAAI: Washington, DC, USA, 2016; Volume 30, pp. 2058–2065. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.