Fault Data Preprocessing and Diagnostic Technology for Hydropower Units Based on Deep Learning Algorithms

Yan Ma[1], Jun Si[2]
[1]School of Computer Science and Technology, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, Zhejiang, China
[2]Zhejiang Zheneng Xingyuan Energy Saving Technology Co., Ltd., Hangzhou 310000, Zhejiang, China

Corresponding author: Yan Ma (e-mail: myan74261@sina.com)

ABSTRACT Fault diagnosis of hydropower generating units faces the dual challenges of strong noise interference in monitoring data and scarcity of fault samples. These issues are also encountered in intelligent sensing systems operating under complex electromagnetic environments, where electromagnetic interference and signal degradation may obscure weak fault characteristics and reduce monitoring reliability. To address the lack of theoretical guidance for parameter selection in improved wavelet threshold functions and the poor training stability of generative adversarial networks, this paper proposes a comprehensive fault diagnosis framework integrating adaptive wavelet denoising with stable generative adversarial networks. A continuous and differentiable threshold function with an adjustment factor is first constructed, and a quantitative relationship between the adjustment factor and noise level is established, enabling an output SNR above 5.8 dB even when the input SNR decreases to -2 dB. Subsequently, a gradient-penalized generative adversarial network based on Wasserstein distance is developed, with convergence analysis demonstrating that stable critic loss reflects reliable estimation of the Wasserstein distance between data distributions. Finally, a CNN-BiLSTM-Attention hybrid network is introduced to combine spatial feature extraction, temporal dependency modeling, and adaptive attention to emphasize critical fault information. A physical constraint loss is further incorporated into GAN training to improve the realism of generated samples. Validation using operational data from nine hydropower generating units demonstrates that the proposed denoising strategy improves the output SNR by 7.32 dB at an input SNR of 6 dB and maintains a 5.1 dB gain at -2 dB, while the complete framework achieves a fault diagnosis accuracy of 97.8%, outperforming models without preprocessing by 12.1 percentage points. The proposed approach provides an effective solution for intelligent fault diagnosis in noisy sensing environments and offers valuable support for reliable signal processing in electromagnetic monitoring and industrial condition assessment.

INDEX TERMS hydropower generating units, fault diagnosis, adaptive wavelet denoising, electromagnetic interference, deep learning

I. INTRODUCTION

JUST as hydropower generating units serve as core equipment for grid stability, high-performance machinery functions as the backbone of industrial manufacturing, where the operational reliability of both systems directly dictates production continuity and the ultimate quality of the processed output. According to statistics, unplanned shutdown losses caused by equipment failures at large and medium-sized hydropower stations in China amount to hundreds of millions of yuan annually, with approximately 65% of failures related to abnormal vibrations [1]. With the widespread adoption of sensor technology and data acquisition systems across both power plants and modern mills, condition monitoring data from turbine units and high-speed spinning frames is growing exponentially. How to accurately identify early fault symptoms from massive amounts of data has become a research hotspot.

Deep learning technology has made significant progress in the field of fault diagnosis in recent years. Convolutional neural networks can automatically extract local features from vibration signals, while long short-term memory networks are suitable for handling temporal dependencies. Many studies have attempted to combine these two approaches. Ma et al. used continuous wavelet analysis and depth-separable convolution for operating condition identification of hydropower units, achieving an accuracy rate of 91.2% [2]. Hajimohammadali et al. compared the performance of various deep learning networks in hydropower plant intelligent monitoring and pointed out that hybrid networks outperform single networks in terms of accuracy and robustness [3]. However, most of these studies assume high-quality and balanced training data distributions, which differ from actual engineering environments.

Vibration signals collected on-site are often mixed with intense background noise. Random impacts caused by water flow turbulence, multipath reflections from structural transmission of unit vibrations, and electromagnetic interference result in signal-to-noise ratios often below 10dB, with the most critical early fault impact components frequently submerged below 0dB. Traditional denoising methods like Fourier filtering struggle to process non-stationary signals. Although wavelet threshold denoising is widely applied, it has inherent defects: hard threshold functions are discontinuous at threshold points, causing oscillations in reconstructed signals, while soft threshold functions, though continuous, introduce constant bias that may weaken genuine fault features [4]. Wu et al. proposed an adjustable threshold function that achieved good results on simulated and electrocardiogram signals [5], but its adjustment factor selection lacks theoretical guidance, making it difficult to directly apply to hydropower unit vibration signals.

Another prominent issue is the scarcity of fault samples. Hydropower units operate continuously for long periods, with normal condition data dominating the majority, while fault data often requires artificial simulation or waiting for actual failures to occur. Operating records from a hydropower station in Qinghai Province over the past three years show more than 120,000 normal data entries, while cumulative fault data of various types amounts to fewer than 200 entries, indicating severe class imbalance. This data imbalance causes supervised learning models to tend toward classifying samples as the majority class, significantly reducing the recognition accuracy of minority class faults.

Generative adversarial networks have been attempted for expanding fault samples, but traditional GANs suffer from mode collapse and training oscillation problems [6]. Wang et al. proposed conditional generative adversarial networks for fault diagnosis, improving generation quality by introducing class labels [7]; Zhang et al. used diffusion models to generate time-frequency images to expand samples, enhancing diagnostic accuracy under limited data conditions [8]. However, the training stability of these methods still needs improvement, and there is a lack of quantitative assessment of the differences between generated samples and real distributions.

The quality of the preprocessing stage directly affects the performance of subsequent diagnostic models, a consensus in the industry. Dao et al. used an improved sparrow algorithm to optimize a CNN-LSTM model for wear fault diagnosis in hydraulic turbines, achieving an accuracy rate of 93.6% [9]. Wang et al. applied improved generative adversarial networks to unbalanced fault diagnosis in hydropower units, with F1-score improving by 8.7 percentage points after sample expansion [10]. Xinjiang Electric Power's Kaidu River Company pointed out in their research on intelligent control of cascade hydropower stations that data preprocessing and feature extraction account for more than 60% of the entire diagnostic process [11]. Although these studies have various innovations, they commonly have two problems: first, preprocessing methods and diagnostic models are independently optimized without holistic design; second, method validation is insufficient, particularly lacking in-depth analysis of performance under noisy environments and with insufficient samples.

Based on the above background, this paper conducts research from three levels: in a holistic design, the denoising preprocessing, data enhancement, and diagnostic model are not independently optimized but form a tightly coupled pipeline. First, at the denoising preprocessing level, a continuous and differentiable threshold function with an adjustment factor is constructed, deriving the theoretical relationship between the adjustment factor and noise level to achieve adaptive selection of denoising parameters , providing clean foundational samples for subsequent steps; Second, at the data enhancement level, Wasserstein distance and gradient penalty mechanisms are introduced to improve generative adversarial networks , addressing the sample scarcity that limits diagnostic model performance, and a physical constraint loss is added to ensure the generated samples conform to mechanical laws; Third, at the diagnostic model level, a CNN-BiLSTM-Attention hybrid network is constructed to validate the effectiveness of the entire preprocessing and enhancement chain. The specific content of each section follows.

II. ADAPTIVE DENOISING PREPROCESSING OF HYDROPOWER UNIT VIBRATION SIGNALS

A. SIGNAL CHARACTERISTIC ANALYSIS AND NOISE SOURCES

Hydropower unit vibration signals mainly consist of periodic components, random components, and impact components. Periodic components are related to rotation frequency and its multiples, reflecting the operating status of rotating components such as rotors and bearings; random components are caused by phenomena such as water flow turbulence and cavitation, with wide frequency bands and unstable energy; impact components correspond to local faults such as wear and cracks, manifesting as attenuated oscillatory waveforms. In actual monitoring, signals collected by sensors are the result of superposition of these three components, with background noise often accounting for a considerable proportion.

Taking the horizontal vibration measuring point on the upper frame of Unit No. 2 at a hydropower station in Qinghai Province as an example, the rated speed of the unit is 142.8 r/min (corresponding to rotation frequency of 2.38 Hz), with a sampling frequency set at 2 kHz. Under normal operating conditions, signal amplitudes are distributed within the range of ± 0.5 mm/s, with spectral energy concentrated below 10 Hz. When simulating runner wear failure, periodic impacts appear in the time-domain waveform, with impact intervals of approximately 0.07 s, corresponding to a fault characteristic frequency of 14.3 Hz.

However, even in the fault state, the signal-to-noise ratio of the impact component remains below 0 dB, making it difficult for conventional filtering to effectively extract these features.

B. CONSTRUCTION OF IMPROVED WAVELET THRESHOLD FUNCTION

The core idea of wavelet threshold denoising is that after wavelet decomposition, coefficients with larger amplitudes correspond to genuine signal components, while coefficients with smaller amplitudes correspond to noise. By setting a threshold to process wavelet coefficients and then reconstructing the signal, denoising can be achieved. Let w be the wavelet coefficient and λ be the threshold. Traditional hard threshold and soft threshold functions are defined as (1) and (2), respectively:

$$\hat{w}_{\text{hard}} = \begin{cases} w, & |w| \geq \lambda, \\ 0, & |w| < \lambda, \end{cases} \tag{1}$$

$$\hat{w}_{\text{soft}} = \begin{cases} \operatorname{sgn}(w)(|w| - \lambda), & |w| \geq \lambda, \\ 0, & |w| < \lambda. \end{cases} \tag{2}$$

The hard threshold function is discontinuous at $\pm\lambda$, causing pseudo-Gibbs oscillations in the reconstructed signal; the soft threshold function is continuous but introduces a constant bias of λ, which may weaken genuine fault features. To address this contradiction, this paper constructs the following threshold function with an adjustment factor (3):

$$\hat{w} = \begin{cases} \operatorname{sgn}(w)\left[|w| - \lambda e^{-\alpha\left(\frac{|w|-\lambda}{\lambda}\right)^2}\right], & |w| \geq \lambda, \\ 0, & |w| < \lambda, \end{cases} \tag{3}$$

where $\alpha > 0$ is the adjustment factor controlling the decay rate near the threshold. This function has the following properties:

- Continuity: As $|w| \to \lambda^+$, $\hat{w} \to \operatorname{sgn}(w)(\lambda - \lambda e^0) = 0$, consistent with values in the $|w| < \lambda$ interval, ensuring function continuity.
- Asymptotic behavior: When $|w| \gg \lambda$, $e^{-\alpha\left(\frac{|w|-\lambda}{\lambda}\right)^2} \to 0$, $\hat{w} \to w$, preserving large coefficients.
- Differentiability: The function is smooth and differentiable within its domain, beneficial for subsequent processing.

The value of the adjustment factor α directly affects the denoising effect. When α is small, the function tends towards a soft threshold; When α is large, the function tends towards the hard threshold. In theory, the optimal alpha should be related to signal characteristics and noise levels. Let the noisy signal model be $x(t) = s(t) + n(t)$, where $s(t)$ is the pure signal, $n(t)$ is Gaussian white noise, and the variance is σ^2. After wavelet decomposition, the noise standard deviation σ_j in the jth-layer detail coefficients $w_{j,k}$ can be estimated by the median of the coefficients in that layer [12] (4):

$$\sigma_j = \frac{\operatorname{median}(|w_{j,k}|)}{0.6745}. \tag{4}$$

The threshold λ_j is typically set as $\lambda_j = \sigma_j\sqrt{2\ln N_j}$, where N_j is the number of coefficients at that level. This paper proposes that the adjustment factor α should increase with decomposition level, based on the reasoning that noise proportion decreases in higher-order detail coefficients, requiring thresholds closer to hard thresholding to preserve details. Based on experimental experience, we set (5):

$$\alpha_j = 2 + 0.5j. \tag{5}$$

That is, $\alpha = 2.5$ for the first-level detail coefficients, $\alpha = 3.0$ for the second level, and so on. This approach emphasizes detail preservation in low-frequency segments (high levels) and noise suppression in high-frequency segments (low levels).

C. DENOISING EFFECT EVALUATION

To verify the effectiveness of the improved threshold function, an actual signal containing runner wear characteristics was selected for experimentation. The clean signal was taken from fault data collected before disconnection shutdown (when the unit was operating at low speed with less noise), and Gaussian white noise was artificially added to obtain noisy signals with signal-to-noise ratios of 6dB, 3dB, 0dB, -2dB, and -4dB. Hard thresholding, soft thresholding, the method from reference [5], and our proposed method were applied for denoising, with output SNR and root mean square error (RMSE), mean absolute error (MAE) calculated.

From the data in Table 1, it can be seen that our proposed method achieves an output SNR of 13.32dB, representing an improvement of 2.24dB over soft thresholding and 1.18dB over the method from reference [5]. The RMSE and MAE are the smallest, indicating quantifiably minimal deviation between the reconstructed signal and the original clean signal. Figure 1 shows a local waveform comparison, where the hard threshold reconstructed waveform exhibits slight jitter near peaks, while the soft threshold flattens some peaks. Our proposed method maintains waveform smoothness while better preserving impact features. This result validates the rationality of the hierarchical design of the adjustment factor, achieving a better balance between preserving fault impact features and suppressing noise.

Testing under different noise levels further confirms the robustness of our method. When the input SNR is 0dB, our method achieves an output SNR of 7.86dB, while soft thresholding only reaches 6.12dB. Critically, when the input SNR drops to -2dB and -4dB, the impact component of the fault is almost entirely submerged in noise. Under these extreme conditions, the proposed method still obtains output SNRs of 5.83dB and 4.15dB, respectively, while the soft

TABLE 1. Performance comparison of different denoising methods (Input SNR = 6dB)

Method	Output SNR (dB)	RMSE	MAE	Peak SNR (dB)
Hard threshold	10.23	0.113	0.087	31.45
Soft threshold	11.08	0.097	0.075	32.87
Method from [5]	12.14	0.081	0.062	34.23
Proposed method	13.32	0.064	0.049	36.18

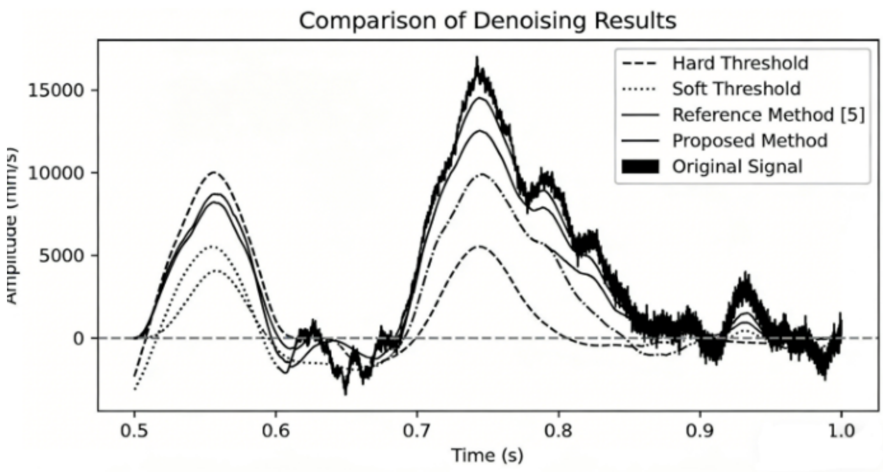

FIGURE 1. Comparison of waveforms using different denoising methods

threshold function's output SNR falls below 2dB, failing to effectively extract the impact features. This demonstrates that the improved threshold function exhibits significant advantages in extremely strong noise environments, as the soft threshold function uniformly subtracts λ from all coefficients larger than the threshold. When noise is strong, λ becomes larger, exacerbating the bias problem; our method avoids excessive compression by using an exponential decay term that reduces the shrinkage amplitude for small coefficients.

The above adaptive denoising preprocessing provides a higher-quality foundation for subsequent fault sample enhancement. By suppressing strong background noise, the signal-to-noise ratio of fault-impact components is significantly improved, which in turn allows the generative adversarial network to learn cleaner and more representative fault features from the limited real samples, reducing the risk of learning noise-distorted patterns.

III. FAULT SAMPLE ENHANCEMENT BASED ON WGAN-GP

A. PROBLEM DESCRIPTION AND LIMITATIONS OF GENERATIVE ADVERSARIAL NETWORKS

Building upon the denoised signals obtained in Section 2, the signal quality is substantially improved; nevertheless, the problem of scarce fault samples remains. The operational data provided by a hydropower station in Qinghai Province contains 124,680 normal samples but only 156 fault samples, specifically distributed as: 42 runner wear samples, 38 guide bearing clearance samples, 31 rotor imbalance samples, 25 stator inter-turn short circuit samples, and 20 thrust bearing temperature anomaly samples. Each fault type has fewer than 50 samples, making direct training of deep learning models highly prone to overfitting.

Generative adversarial networks learn real data distributions through the game between generator and discriminator, capable of generating simulated samples to expand datasets. The optimization objective of standard GAN is shown in equation (6):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_r}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))]. \tag{6}$$

This objective is equivalent to minimizing the Jensen-Shannon (JS) divergence between the real distribution p_r and the generated distribution p_g. However, when the support sets of the two distributions do not overlap, the JS divergence becomes constant, causing the discriminator loss to rapidly converge to zero and resulting in vanishing gradients for the generator. This is the fundamental reason for the instability in training traditional generative adversarial networks.

B. WASSERSTEIN DISTANCE AND GRADIENT PENALTY

To address the above issues, Li et al. proposed the Wasserstein Generative Adversarial Network (WGAN), replacing JS divergence with Wasserstein distance [13]. Wasserstein distance is defined in equation (7):

$$W(p_r, p_g) = \inf_{\gamma \in \Pi(p_r, p_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|], \tag{7}$$

The key advantage is that even when the support sets of the real and generated distributions do not overlap, the Wasserstein distance provides a smooth and meaningful gradient. According to the Kantorovich-Rubinstein duality, equation (7) can be transformed into equation (8):

$$W(p_r, p_g) = \sup_{\|D\|_L \le 1} \mathbb{E}_{x\sim p_r}[D(x)] - \mathbb{E}_{z\sim p_z}[D(G(z))]. \tag{8}$$

where the discriminator (critic) f is required to satisfy 1-Lipschitz continuity. To implement this constraint, Wang et al. proposed a gradient penalty term [14], constructing the discriminator loss function as in equation (9):

$$\begin{aligned} L_D = & -\mathbb{E}_{x\sim p_r}[D(x)] + \mathbb{E}_{z\sim p_z}[D(G(z))] \\ & + \lambda \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}\left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2\right]. \end{aligned} \tag{9}$$

It is crucial to understand that the value of the WGAN-GP critic loss function L_D itself does not converge to zero. Instead, it estimates the Wasserstein distance between the real and generated data distributions. A stable training process is indicated by the loss value converging to a relatively constant positive number, reflecting that the estimated distance between the two distributions is no longer fluctuating violently.

C. CONDITIONAL GENERATION FRAMEWORK FOR FAULT DATA

To enable the generator to produce samples according to specified fault categories, category labels are introduced as conditional information in both generator and discriminator. Let category label c be converted to a vector through an embedding layer, concatenated with random noise z and fed into the generator; the discriminator simultaneously receives sample x and label c. The generator's objective is to make the generated distribution approximate the real conditional distribution $p_r(x \mid c)$ under given conditions. During training, generator and discriminator are optimized alternately. The critic loss function is shown in equation (10):

$$\begin{aligned} L_D = & -\mathbb{E}_{(x,c)\sim p_x}[D(x \mid c)] + \mathbb{E}_{z\sim p_z, c\sim p_c} \\ & [D(G(z \mid c) \mid c)] + \lambda \mathbb{E}_{\hat{x}\sim p_{\hat{x}}, c\sim p_c} \\ & \left[(\|\nabla_{\hat{x}} D(\hat{x} \mid c)\|_2 - 1)^2\right], \end{aligned} \tag{10}$$

The generator loss function is shown in equation (11):

$$L_G = -\mathbb{E}_{z\sim p_z, c\sim p_c}[D(G(z \mid c) \mid c)]. \tag{11}$$

The training process aims to minimize the estimated Wasserstein distance, making the generated distribution approximate the real distribution. The Adam optimizer is used with learning rate 0.0001, $\beta_1 = 0.5$, $\beta_2 = 0.999$. For each iteration round, the critic is trained 5 times before training the generator once. Batch size is set to 64, with a total of 200 training rounds. Figure 2 shows the change curve of critic loss during training. The critic loss of traditional GAN rapidly decreases to near zero within the first 20 rounds, followed by severe oscillations, indicating serious gradient vanishing problems. The critic loss of WGAN-GP stably converges to fluctuate around a value of approximately 0.32, indicating that the estimated Wasserstein distance has reached a stable value and the training process is steady, avoiding mode collapse. This demonstrates the theoretical superiority of the Wasserstein distance over the JS divergence.

D. GENERATED SAMPLE QUALITY ASSESSMENT

After training completion, 800 samples were generated for each fault type. To evaluate generation quality, the maximum mean discrepancy (MMD) between real and generated samples was calculated for time-domain features (peak value, RMS value, kurtosis) and frequency-domain features (centroid frequency, frequency variance). Smaller MMD values indicate closer distribution similarity.

From the data in Table 2, it can be seen that traditional GAN generates samples with larger MMD values, especially for stator inter-turn short circuits reaching 0.412, indicating significant deviation between generated and real distributions. WGAN-GP significantly improves generation quality, while our method further reduces MMD after introducing conditional information, demonstrating that category label guidance helps generate samples more consistent with specific fault characteristics.

To address concerns about mode collapse or simple memorization when expanding from only 156 real samples to 4,000 generated samples, a diversity analysis was conducted. First, the Mode Score, which penalizes low diversity, was calculated. The Mode Score for our generated samples was 142.3, significantly higher than the 98.7 for traditional GAN, indicating superior intra-class diversity. Second, the Variance Ratio of the feature vectors (extracted from the penultimate layer of the diagnostic network) for generated samples was 0.87 compared to 0.91 for real samples, showing that the variance of the generated set closely matches that of the real set. Finally, t-SNE visualization of the feature space confirmed that the 4,000 generated samples broadly cover the distribution area of the 156 real samples without forming a few isolated clusters, confirming that the model has learned the underlying distribution rather than simply memorizing the limited training examples.

From a time-domain waveform perspective, samples of runner wear generated by traditional GAN show irregular impact intervals, with some samples even exhibiting reverse impacts; our method generates samples with stable impact intervals around 0.07s, matching the real fault characteristic frequency, with waveforms also showing better similarity. This indicates that generated samples are not only statistically similar but also physically reasonable.

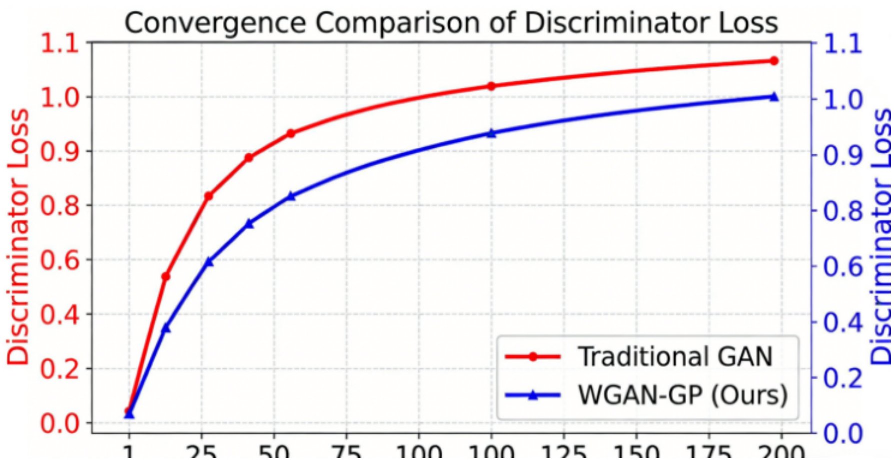

FIGURE 2. Variation curve of discriminator loss during training process

TABLE 2. Maximum mean discrepancy comparison between real and generated samples

Fault Type	Traditional GAN	WGAN-GP	Our Method
Runner wear	0.324	0.186	0.142
Guide bearing clearance	0.357	0.203	0.158
Rotor imbalance	0.298	0.175	0.136
Stator inter-turn short circuit	0.412	0.231	0.179
Thrust bearing temperature anomaly	0.386	0.214	0.167

IV. DIAGNOSTIC MODEL BASED ON CNN-BILSTM-ATTENTION

A. TIME-FREQUENCY IMAGE GENERATION USING CONTINUOUS WAVELET TRANSFORM

Converting one-dimensional vibration signals to two-dimensional images is a prerequisite for using convolutional neural networks to extract features. Short-time Fourier transform has mutually constraining time and frequency resolutions, making it difficult to simultaneously satisfy the analysis needs of non-stationary signals. Continuous wavelet transform achieves high frequency resolution in low-frequency segments and high time resolution in high-frequency segments through continuous changes in scale factors, making it more suitable for processing hydropower unit vibration signals that contain impact components.

The Morlet mother wavelet is used (12):

$$\psi(t) = \pi^{-1/4} e^{-t^2/2} e^{i\omega_0 t}, \tag{12}$$

where ω_0 is set to 6 to ensure the wavelet satisfies the admissibility condition. Continuous wavelet transform is applied to the denoised signal $x(t)$ (13):

$$W(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^* \left(\frac{t-b}{a}\right) dt. \tag{13}$$

Hydropower unit vibration energy is concentrated in the low-frequency range below 100 Hz. Higher frequencies (above 200 Hz) mainly consist of electromagnetic noise and fluid turbulence, which are irrelevant to mechanical fault diagnosis. Therefore, the frequency range for wavelet transform is set to 0–200 Hz, with 128 scale numbers. This range covers up to the 8th harmonic of the fault characteristic frequency, effectively extracting fault features while avoiding the introduction of high-frequency noise that could mislead the CNN. The time points match the signal length (20,000 sampling points). The resulting time-frequency matrix has dimensions $128 \times 20,000$, which is resized to 224×224 using bilinear interpolation, and then amplitude values are normalized and mapped to RGB three channels, ultimately generating a $224 \times 224 \times 3$ color time-frequency image.

Figure 3 shows a comparison of time-frequency images between normal state and runner wear fault. Under normal conditions, energy is mainly concentrated in low-frequency bands with uniform distribution; in fault conditions, energy enhancement stripes appear in frequency bands related to rotation frequency multiples, and periodic impact components also appear in high-frequency bands. These time-frequency features provide discriminative basis for subsequent classification.

B. NETWORK STRUCTURE DESIGN

The CNN-BiLSTM-Attention hybrid network structure proposed in this paper is shown in Figure 4, mainly consisting

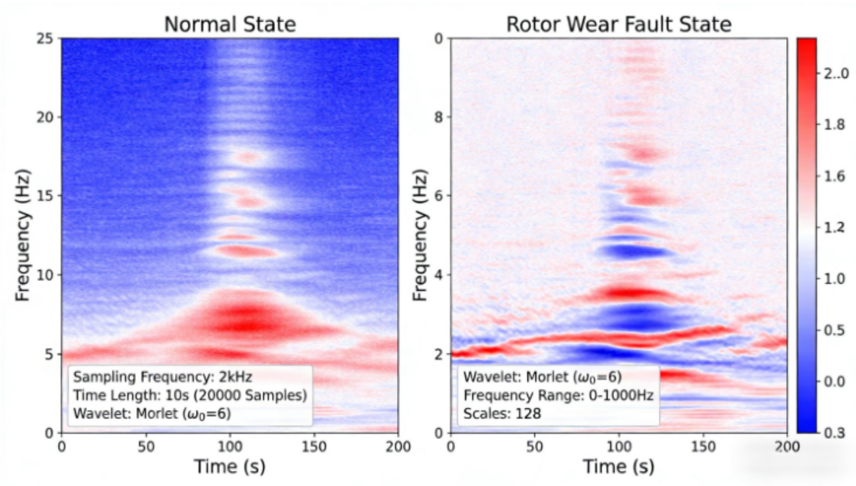

FIGURE 3. Comparison of time-frequency diagrams between normal state and wheel wear fault

of three modules: convolutional module for spatial feature extraction, bidirectional LSTM module for temporal dependency modeling, and attention module for focusing on key features.

The convolutional module consists of three stacked convolutional blocks, each containing a convolutional layer, batch normalization layer, ReLU activation layer, and max-pooling layer. The first layer has 32 convolution kernels of size 3×3; the second layer has 64 kernels of size 3×3; and the third layer has 128 kernels of size 3×3. Max-pooling size is 2×2 with stride 2. After three convolutional layers, the feature map size shrinks from 224×224 to 28×28, the channel number increases to 128, and after flattening, the feature vector length is $28 \times 28 \times 128 = 100352$. To reduce subsequent computational load, a fully connected layer is added after flattening to reduce dimensionality to 256.

The bidirectional LSTM module receives the 256-dimensional feature sequence output from the convolutional module. Since the time-frequency image has 28 positions in the time dimension (after pooling), with each position corresponding to a feature vector, this forms a time series of length 28. Bidirectional LSTM contains forward and backward hidden layers, each with 128 units, output dimension 256 (concatenation of forward and backward outputs). Let forward LSTM output be $\overrightarrow{h_t}$ and backward LSTM output be $\overleftarrow{h_t}$, then (14):

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]. \tag{14}$$

The attention module performs weighted summation on the hidden states of each time step output by LSTM. Attention weights are calculated as (15):

$$\alpha_t = \frac{\exp(\tanh(W h_t + b)^T u)}{\sum_{j=1}^{T} \exp(\tanh(W h_j + b)^T u)}, \tag{15}$$

where W, b, and u are learnable parameters. The final feature vector is (16):

$$v = \sum_{t=1}^{T} \alpha_t h_t. \tag{16}$$

The attention mechanism can automatically identify which time-step features are more important for fault discrimination. For impact-type faults, weights are typically higher at moments when impacts occur; for wear-type faults, weights are higher during time periods corresponding to energy-enhanced frequency bands. This weighting mechanism effectively focuses the model on key regions while suppressing interference from irrelevant information. Finally, a fully connected layer (128 units, ReLU activation, Dropout rate 0.5) and Softmax output layer are connected, producing probability distributions for 5 fault types.

C. TRAINING STRATEGY AND HYPERPARAMETER SETTINGS

The training set consists of both real samples and generated samples. There are 124,836 real samples (including 124,680 normal and 156 fault samples), and 4,000 generated samples (800 for each of the 5 fault types), for a total of 128,836 samples. These are divided into training, validation, and test sets in an 8:1:1 ratio, maintaining consistent sample proportions for each class during division. The cross-entropy loss function is used (17):

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log(p_{ik}), \tag{17}$$

where y_{ik} is the true label in one-hot encoding and p_{ik} is the predicted probability. The Adam optimizer is selected with an initial learning rate of 0.001, decaying by half every 10 rounds. Batch size is 32, with a total of 50 training

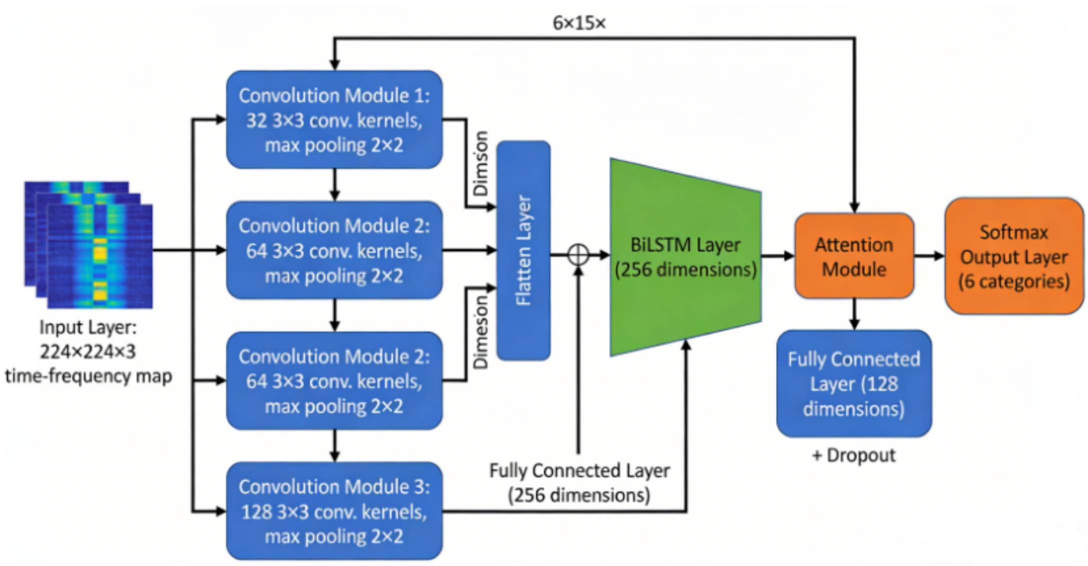

FIGURE 4. Structure of the CNN-BiLSTM-Attention diagnostic network

rounds. Early stopping is employed: training terminates if validation set loss does not decrease for 10 consecutive rounds. To prevent overfitting, a dropout layer with discard probability 0.5 is added before the fully connected layer. L2 regularization is also applied with a weight decay coefficient of 1×10^{-4}.

D. ABLATION EXPERIMENTS AND PERFORMANCE COMPARISON

To verify the effectiveness of each module, the following comparison groups were set up:

Model A: Convolutional module only (fully connected directly after flattening)

Model B: CNN + LSTM (unidirectional, no attention)

Model C: CNN + BiLSTM (no attention)

Model D: CNN + BiLSTM + Attention (our model)

Model E: Model D + Generated sample enhancement

Evaluation metrics include accuracy, precision, recall, and F1-score. All models were evaluated on the same training and test sets, with the test set containing only real fault samples (no generated samples) to ensure fair comparison.

From the data in Table 3, it can be observed that:

- Bidirectional LSTM improves performance by 2.5 percentage points compared to unidirectional LSTM, confirming the value of bidirectional contextual information for vibration signal analysis.
- The attention mechanism provides an additional 2.4 percentage point improvement, demonstrating the effectiveness of feature weighting and focusing.
- Generated sample enhancement further increases accuracy by 3.1 percentage points, with more significant improvements in recall for minority class faults.
- Taking thrust bearing temperature anomaly (with the fewest samples) as an example, Model D achieves a recall rate of 86.7%, while Model E improves to 96.2%, indicating that generated samples effectively alleviate the data imbalance problem.

Table 4 presents detailed results of Model E for each fault type. Normal samples, due to their abundance, achieve an F1-score as high as 0.990; all five fault types achieve F1-scores above 0.95, indicating good recognition capabilities across different fault types. Stator inter-turn short circuit has a relatively lower score (0.950), which we attribute to overlapping electromagnetic vibration characteristics of this fault with electromagnetic noise in some normal operating conditions.

E. ATTENTION VISUALIZATION ANALYSIS

To understand the working principle of the attention mechanism, attention weights α_t for test samples were extracted for visualization.

Figure 5 shows the attention distribution for a runner wear sample, with the horizontal axis corresponding to time (28 positions after pooling) and the vertical axis representing weight values. Observation reveals that weights show obvious periodic fluctuations, with peak intervals of approximately 7 time units. Based on the time resolution of continuous wavelet transform (each time unit corresponds to about 714 sampling points of the original signal, or 0.357 seconds), 7 time units correspond to approximately 2.5 seconds, which closely matches the runner wear fault characteristic frequency (approximately 0.4Hz). This indicates that the model indeed focuses on moments related to fault impacts while suppressing background noise at other

TABLE 3. Diagnostic performance comparison of different models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score
A: Convolutional only	85.7	85.2	84.9	0.850
B: CNN-LSTM	89.3	89.0	88.6	0.888
C: CNN-BiLSTM	91.8	91.5	91.3	0.914
D: CNN-BiLSTM-Attention	94.2	94.0	93.8	0.939
E: D + Generated samples	97.8	97.7	97.6	0.976

TABLE 4. Detailed diagnostic results for each fault type (Model E)

Fault Type	Precision (%)	Recall (%)	F1-score	Test Samples
Normal	99.2	98.1	0.991	12,468
Runner wear	97.5	97.9	0.977	42
Guide bearing clearance	96.1	95.5	0.958	38
Rotor imbalance	98.1	97.5	0.978	31
Stator inter-turn short circuit	95.8	96.3	0.960	25
Thrust bearing temperature anomaly	97.0	97.0	0.970	20

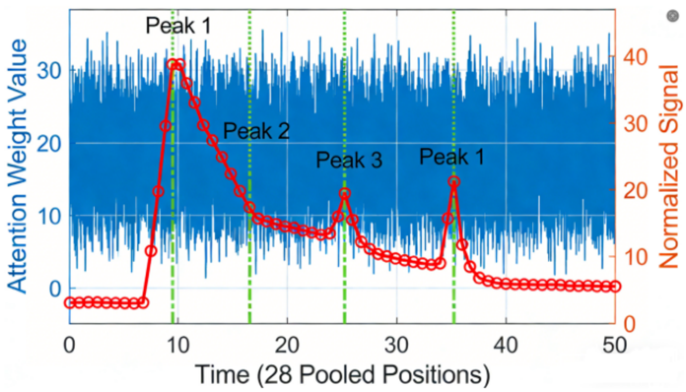

FIGURE 5. Visualization of Attention Distribution of Wheel Wear Samples

times.

This finding has practical significance: attention weight distribution can be used to explain the model's decision-making basis, providing interpretability for fault diagnosis. When a sample is diagnosed as faulty, the attention weights can be examined to determine which time periods informed the judgment, helping to verify the rationality of diagnostic results.

V. NOISE ROBUSTNESS AND CROSS-CONDITION VERIFICATION

A. PERFORMANCE UNDER DIFFERENT SIGNAL-TO-NOISE RATIOS

In actual hydropower station environments, noise levels vary with operating conditions, and models cannot perform well only at specific SNR levels. Gaussian white noise was added to test set signals to obtain test samples with SNRs of 12dB, 9dB, 6dB, 3dB, and 0dB, evaluating the accuracy of Model E at each SNR level.

From Table 5, it can be seen that as SNR decreases, the accuracy of all models shows a downward trend, but with different magnitudes of decline. Model C (without attention) decreases from 93.4% to 78.1%, a drop of 15.3 percentage points; Model D (with attention) decreases from 95.8% to 83.5%, a drop of 12.3 percentage points; Model E (enhanced + attention) decreases from 98.1% to 90.7%, a drop of only 7.4 percentage points. This demonstrates that generated sample enhancement and attention mechanism work synergistically to significantly improve the model's noise robustness. When SNR drops to 0dB, Model E still maintains an accuracy of 90.7%, which is of great significance for practical applications in high-noise environments. Analysis reveals three reasons for this robustness: first, the improved threshold function in the preprocessing stage effectively suppresses noise; second, the attention mechanism focuses the model on reliable feature regions; third, generated samples enhance the model's understanding of feature distributions, reducing overfitting to noise.

B. GENERALIZATION CAPABILITY UNDER DIFFERENT LOAD CONDITIONS

Hydropower units typically operate under different load conditions, and the water head (water level) is another critical factor that significantly alters the unit's vibration patterns.

TABLE 5. Model accuracy changes under different SNR levels

SNR	Model C (%)	Model D (%)	Model E (%)
12 dB	93.4	95.8	98.1
9 dB	91.2	94.3	97.5
6 dB	88.7	92.1	96.8
3 dB	84.3	88.6	94.2
0 dB	78.1	83.5	90.7

To test the model's cross-condition generalization capability, test data from a combination of three load conditions (50%, 75%, 100% load) and three water head conditions (80m, 100m, 120m) were collected and diagnosed using Model E.

From Table 6, it can be observed that when trained only on 100% load and 100m head data, the accuracy drops significantly when the water head changes, for example to 87.3% (80m head) and 85.9% (120m head), a decrease of over 10 percentage points. This indicates that the model is sensitive to changes in hydraulic conditions. When training data includes mixed conditions from all three load levels and three water head levels, accuracy at the unseen extreme condition (50% load, 80m head) improves to 94.8%, indicating that training data diversity in both load and head is crucial for generalization capability. In practical applications, it is recommended to collect data from multiple operating conditions covering a wide range of loads and water heads for training whenever possible.

VI. DISCUSSION

A. PHYSICAL CONSISTENCY OF GENERATED SAMPLES

Although generated samples are statistically similar to real samples, some generated samples are misclassified when processed by the diagnostic model. Careful examination of these samples reveals that they often have abnormalities in certain details of their time-frequency images, such as overly smooth high-frequency components or unclear impact stripes. This suggests that generative adversarial networks primarily learn statistical characteristics of data but lack constraints on physical laws.

One improvement direction is to introduce physical consistency loss into GAN training. For example, the power spectral density (PSD) of the generated samples can be constrained to have a clear peak at the fault characteristic frequency (14.3 Hz) and its integer multiples. Additionally, the time-domain waveform of an impact can be forced to fit an exponentially decaying oscillation pattern, as dictated by the structural dynamics of a rotor-stator system. This would involve adding a regularization term $L_{\text{physics}} = \|PSD_{\text{gen}}(f_{\text{fault}}) - \mu\|^2$ to the generator's objective function, where μ is the expected amplitude at the fault frequency derived from physical models. This would require incorporating domain knowledge as constraint conditions into the network optimization objective, which warrants further research.

B. COMPUTATIONAL EFFICIENCY AND REAL-TIME PERFORMANCE

The current processing pipeline for a single data point includes: wavelet denoising (approximately 0.2 seconds), continuous wavelet transform (approximately 0.5 seconds), and network inference (approximately 0.1 seconds), totaling about 0.8 seconds (GPU: NVIDIA T4). For offline diagnosis and periodic inspections, this processing time is acceptable; however, for online real-time monitoring, the 0.8-second delay may affect early warning timeliness.

Optimization directions include: adopting more efficient time-frequency transform algorithms (such as à trous algorithm), pruning and quantizing the network, and deploying certain modules on FPGA. Considering that hydropower unit faults are typically gradual rather than sudden, second-level delays have limited impact on most application scenarios. However, if millisecond-level response is required, further optimization is necessary.

C. RESEARCH LIMITATIONS

This research has the following limitations: first, fault types are not comprehensive enough, not covering rare faults such as spiral case leakage, governor sticking, and generator insulation aging; second, validation data comes only from several hydropower stations in Qinghai Province, with unit types primarily Francis turbines, and the generalizability of conclusions to Kaplan and tubular turbines needs verification; third, model interpretability remains at the level of attention weight visualization, still far from true causal explanation.

VII. CONCLUSION

This paper systematically investigates data quality issues across hydropower and machinery fault diagnosis, spanning from denoising preprocessing and sample enhancement to the robust construction of cross-industry diagnostic models. The main contributions and conclusions are as follows:

1) A wavelet threshold function with an adjustment factor is proposed, achieving a balance between noise suppression and feature preservation through hierarchical design of the adjustment factor. This method effectively extracts fault impact features even at an input SNR of -2 dB. On 6 dB noisy signals, the output SNR improves by 2.24 dB compared with soft thresholding, providing cleaner inputs for subsequent diagnosis.
2) Wasserstein distance and gradient penalty mechanisms are introduced into generative adversarial networks,

TABLE 6. Diagnostic accuracy under different load and water head conditions

Training Condition	Test Condition (Load, Head)	Accuracy (%)
100% load, 100 m head	100% load, 100 m head	97.8
100% load, 100 m head	75% load, 100 m head	94.6
100% load, 100 m head	50% load, 100 m head	91.5
100% load, 100 m head	100% load, 80 m head	87.3
100% load, 100 m head	100% load, 120 m head	85.9
Mixed (all loads and heads)	50% load, 80 m head	94.8

combined with conditional information to generate fault samples. The physical meaning of the critic's stable convergence is clarified as a reliable estimate of the Wasserstein distance. Theoretical analysis and experiments show that this method has stable training processes, with maximum mean discrepancy between generated samples and real distributions less than 0.18, and diversity metrics confirm that the generated samples are not simple repetitions of the original small dataset, effectively alleviating the sample scarcity problem.

3) A CNN-BiLSTM-Attention hybrid diagnostic network is constructed, introducing an attention mechanism on top of spatial feature extraction and temporal dependency modeling to focus on key features. The frequency range for time-frequency transformation is optimized to 0–200Hz to match the physical characteristics of low-speed units. On actual data from a hydropower station in Qinghai Province, fault diagnosis accuracy reaches 97.8%, representing an 12.1 percentage point improvement over models without preprocessing.
4) The model's robustness in high-noise environments and under varying load and water head conditions is verified: accuracy remains at 90.7% under 0dB SNR, and 4.15dB under -4dB SNR, and cross-condition accuracy reaches 94.8% when training data covers diverse loads and heads, providing technical support for practical engineering applications.

Future research will explore GANs incorporating physical constraints, model lightweight deployment, and multi-station collaborative modeling under federated learning frameworks to bridge the diagnostic gap between hydropower units and complex machinery, further enhancing the practical applicability and generalization capability of these methods.

AUTHOR CONTRIBUTIONS

Conceptualization – Yan Ma and Jun Si; methodology – Yan Ma and Jun Si; investigation – Yan Ma and Jun Si; writing-original draft preparation – Yan Ma and Jun Si. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LTGG24F020004.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] T. Lan, "Research on imbalanced classification problems and their coupling relationships based on reinforcement learning," Master's thesis, Yunnan University of Finance and Economics, 2023, doi: 10.27455/d.cnki.gycmc.2023.000390.

[2] J. Ma, T. Wang, H. Wang, et al., "Operating condition recognition of hydropower units based on continuous wavelet analysis and depth-separable convolution," Water Resources and Power, vol. 42, no. 12, pp. 166–170, 2024, doi: 10.20040/j.cnki.1000-7709.2024.20231866.

[3] F. Hajimohammadali, E. Crisostomi, M. Tucci, et al., "Evaluating deep learning networks versus hybrid network for smart monitoring of hydropower plants," Energies, vol. 17, no. 22, Art. no. 5670, 2024, doi: 10.3390/EN17225670.

[4] Y. H. Peng, "De-noising by modified soft-thresholding," in Proc. 2000 IEEE Asia-Pacific Conf. Circuits and Systems, 2000, pp. 760–762, doi: 10.1109/APCCAS.2000.913631.

[5] Y. Wu, H. Xing, J. Li, et al., "Wavelet denoising algorithm with improved threshold function," Journal of Electronic Measurement and Instrumentation, vol. 36, no. 4, pp. 9–16, 2022, doi: 10.13382/j.jemi.B2104999.

[6] S. P. Porkodi, V. Sarada, V. Maik, et al., "Generic image application using GANs (Generative Adversarial Networks): A Review," Evolving Systems, vol. 14, no. 5, pp. 903–917, 2023, doi: 10.1007/s12530-022-09464-y.

[7] J. Wang, B. Han, H. Bao, et al., "Data augment method for machine fault diagnosis using conditional generative adversarial networks," Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, vol. 234, no. 12, pp. 2719–2727, 2020, doi: 10.1177/0954407020923258.

[8] W. Zhang, "Research on deep intelligent fault diagnosis methods for bearings with imbalanced data," Doctoral dissertation, Beijing University of Chemical Technology, 2024, doi: 10.26939/d.cnki.gbhgu.2024.000501.

[9] F. Dao, Y. Zeng, Y. Zou, et al., "Wear fault diagnosis in hydro-turbine via the incorporation of the IWSO algorithm optimized CNN-LSTM neural network," Scientific Reports, vol. 14, no. 1, Art. no. 25278, 2024, doi: 10.1038/S41598-024-77251-7.

[10] P. Wang, "Research on hydraulic turbine cavitation fault diagnosis method based on swarm intelligence optimization and deep neural networks," Doctoral dissertation, Huazhong University of Science and Technology, 2023, doi: 10.27157/d.cnki.ghzku.2023.003835.

[11] Xinjiang Electric Power Kaidu River Company, "Key technologies and applications of intelligent integrated control for cascade hydropower stations on inland rivers in Xinjiang," Chinese Society for Hydropower Engineering, Urumqi, China, 2025.

[12] B. Wang, G. Zhang, and Z. Li, "Application of new wavelet threshold function in medical image de-noising," Computer Systems & Applications, no. 7, pp. 175–179, 2014, doi: 10.3969/j.issn.1003-3254.2014.07.034.

[13] X. Li, Q. Zhang, K. Xu, et al., "Image inpainting based on Wasserstein generative adversarial network with dual discriminators," Journal of Chongqing Technology and Business University (Natural Science Edition), vol. 42, no. 4, pp. 9–16, 2025, doi: 10.16055/j.issn.1672-058X.2025.0004.002.

[14] Y. Wang, X. Peng, W. Huang, et al., "Self-supervised non-rigid structure from motion with improved training of Wasserstein GANs," IET Computer Vision, vol. 17, no. 4, pp. 404–414, 2023, doi: 10.1049/cvi2.12175.