

Article

Lightweight Vision Transformer for Real-Time Threat Level Assessment in Φ -OTDR-Based Pipeline Monitoring

Yuhan Zhang ¹, Hao Zeng ^{2,3}, Chang Su ¹, Jie Yang ¹, Jianjun Zhu ^{4,*} and Jianli Wang ^{1,*}

¹ Jiangsu Key Laboratory for Design and Manufacturing of Precision Medicine Equipment, Department of Mechanical Engineering, Southeast University, Nanjing 210096, China; 220245147@seu.edu.cn (Y.Z.); 220245164@seu.edu.cn (C.S.); 220230456@seu.edu.cn (J.Y.)

² State Key Laboratory of Shale Oil and Gas Enrichment Mechanisms and Effective Development, Beijing 100083, China; zenghao.syky@sinopec.com

³ State Center for Research and Development of Oil Shale Exploitation, Beijing 100083, China

⁴ College of Mechanical and Transportation Engineering, China University of Petroleum, Beijing 102249, China

* Correspondence: jianjun-zhu@cup.edu.cn (J.Z.); wangjianli@seu.edu.cn (J.W.)

Abstract

Phase-sensitive optical time domain reflectometry (Φ -OTDR) is a highly sensitive distributed vibration sensing technology crucial for pipeline safety monitoring. However, its sensitivity makes it susceptible to environmental interference, leading to frequent false alarms by misclassifying routine activities as threats. To enable accurate threat identification and rapid response, this study proposes a lightweight LightPatch Vision Transformer (LP-ViT) model suitable for edge deployment. We establish a mapping between excavator-pipeline distance and threat levels: “direct intrusion” (within 5 m), “high-risk operation” (within 10 m), and “background construction” (beyond 15 m). The LP-ViT model is developed through structural optimization and parameter compression of the standard Vision Transformer, achieving a 96.6% reduction in parameter count while maintaining a high classification accuracy of 89.9%. Furthermore, via knowledge distillation, we derive an ultra-lightweight student model with merely 0.37 M parameters, which achieves an inference latency of 5.5 ms per sample, enabling millisecond-level threat detection and response. The proposed solution effectively enhances both the classification accuracy and real-time performance of Φ -OTDR systems in complex environments, providing a practical pathway for implementing edge intelligence in pipeline safety monitoring.

Keywords: distributed optical fiber sensing; knowledge distillation; lightweight model; pipeline integrity; threat detection; transformers; Φ -OTDR



Academic Editor: Ephraim Suhir

Received: 29 December 2025

Revised: 28 January 2026

Accepted: 31 January 2026

Published: 6 February 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Linear infrastructure such as pipelines constitutes a typical class of critical assets, whose safe operation is closely associated with energy delivery assurance, public safety, and environmental risk control [1]. To meet the safety-governance requirements across the pipeline lifecycle, risk-informed asset integrity management (AIM) frameworks have attracted increasing attention; these frameworks emphasize the coordinated integration of online monitoring, condition assessment, and maintenance decision-making [2] to improve the timeliness and effectiveness of risk identification and mitigation. Within this context, the value of monitoring systems extends beyond anomaly detection: under conditions of high alarm volumes, complex background disturbances, and constrained response resources, monitoring outputs are expected to provide actionable risk stratification and/or

response-priority information, thereby enabling rapid decision-making for patrol planning and emergency response [3].

Fiber-optic intrusion detection systems integrate optical sensing, electronics, and signal processing, and have become a key technological route for long-range security monitoring [4]. Among them, phase-sensitive optical time-domain reflectometry (Φ -OTDR) is a highly sensitive distributed vibration sensing technique that enables long-distance, continuous disturbance detection along an optical fiber with relatively high spatial resolution [5]. Consequently, it has been widely adopted in applications such as oil and gas pipelines, railways, and perimeter security [6–8]. Existing Φ -OTDR studies have primarily focused on disturbance detection and event-type recognition, where time-/frequency-domain features or deep networks are used to detect, classify, and localize intrusion events, thereby improving recognition accuracy and noise robustness.

However, in real-world operation and maintenance scenarios, monitoring systems often face dense alarms and limited response resources, and outputs limited to “event categories” are insufficient to directly support risk stratification and response-priority decisions in patrol and emergency workflows. Therefore, how to achieve both disposal-oriented threat-level outputs and edge-deployable low-latency inference within the Φ -OTDR framework remains an open problem and deserves further investigation.

Under high-sensitivity distributed vibration monitoring, Φ -OTDR systems are susceptible to complex background disturbances, which can degrade the signal-to-noise ratio (SNR) of the acquired signals and induce feature overlap. Factors such as airflow, vibrations from operating equipment, and environmental scattering may produce time- and frequency-domain responses that resemble those of intrusion events, thereby increasing false-alarm rates and reducing event separability [9]. When alarm density is high, a large number of low-risk or invalid alarms can consume response resources and trigger “alarm fatigue,” undermining operators’ trust in warnings and their response efficiency [10]. Moreover, frequent false alarms may obscure early indicators of high-risk intrusions at the information level, leading to delayed responses and elevated safety risks. Therefore, for engineering applications, Φ -OTDR monitoring should go beyond outputs, such as “whether an event occurs” or “event category,” and instead provide threat grading and response-priority information that can directly support patrol and emergency workflows, enabling the prioritization of high-risk warnings under resource-constrained conditions [11].

To reduce false alarms and improve event discriminability, the Φ -OTDR community has extensively investigated machine-learning-based recognition methods [12,13]. Early studies typically relied on handcrafted features and conventional classifiers [14–16]; however, their generalization is often limited under strong noise and varying operating conditions, and they usually involve lengthy preprocessing and feature-engineering pipelines with substantial dependence on expert experience. In recent years, deep learning has gradually become the dominant paradigm [17]. Methods, such as 1D/2D convolutional neural networks (CNNs), recurrent neural networks (RNNs)/long short-term memory (LSTM) architectures, and hybrid variants, have achieved notable improvements in recognition accuracy [18–23].

Nevertheless, from the perspectives of engineering deployment and closed-loop risk management, existing studies still exhibit three main limitations. (1) Most works focus on event detection and/or event-type identification as the primary outputs. Only a few directly provide threat levels or risk priorities, creating a disconnect between monitoring results and patrol/emergency decision-making [24]. (2) Under complex noise backgrounds, Φ -OTDR signals often exhibit correlated structures across time slices and long-range dependencies. Models that rely mainly on convolutional features with local receptive fields, or sequence modeling dominated by short-term dynamics, may fail to robustly capture global

context, thereby remaining constrained in fine-grained discrimination. (3) Deep networks adopted to pursue higher accuracy typically introduce substantial parameter counts and computational overhead, which makes low-latency, resource-efficient online inference on edge devices challenging. Therefore, it is necessary to explore lightweight model frameworks that simultaneously offer global dependency modeling and edge deployability to support disposal-oriented threat-level outputs.

In this work, we propose a lightweight Vision Transformer framework for online Φ -OTDR monitoring, termed the LightPatch Vision Transformer (LP-ViT), to enable disposal-oriented fine-grained threat-level assessment. We use the relative distance between intrusive activities and the pipeline as an actionable proxy variable for response risk, and map alarms to multiple threat levels. This design more directly supports priority decisions in on-site patrol and emergency response, while reducing unnecessary dispatches triggered by low-risk activities. From a modeling perspective, LP-ViT leverages self-attention to enhance the representation of global context and long-range dependencies embedded in time–frequency textures, while adopting a lightweight architecture to reduce parameter size and computational cost. Furthermore, we incorporate knowledge distillation to transfer discriminative representations from a high-capacity teacher model to a compact student model, thereby meeting the low-latency inference requirements of edge deployment without sacrificing recognition performance. In the experiments, we systematically evaluate the proposed approach from both recognition effectiveness and deployment cost, and analyze the accuracy–latency trade-off and its practical engineering applicability. Despite these advances, this study has limitations: the validation is conducted on data collected from the current set of scenarios, and cross-scenario and cross-operating-condition generalization remains to be further assessed under larger-scale and more diverse datasets. Overall, our central objective is to unify “disposal-oriented threat-level outputs” with “edge-deployable real-time inference” in Φ -OTDR monitoring.

The remainder of this paper is organized as follows. Section 2 presents the methodology, including data preprocessing and time–frequency image construction, the lightweight architecture design of LP-ViT, and the knowledge distillation strategy. Section 3 describes the experimental data and evaluation protocol, reports comparative results against baseline methods and recent lightweight models tailored for edge deployment in Φ -OTDR applications, and further presents edge-side inference costs, as well as ablation studies and sensitivity analyses of key settings. Section 4 discusses the experimental findings, summarizes the main contributions and engineering implications, and outlines the study’s limitations and directions for future work.

2. Methodology

2.1. Principles of Φ -OTDR Distributed Vibration

The distributed vibration sensing system developed in this study is based on Φ -OTDR technology. The hardware configuration and sensing principle are shown in Figure 1. The system consists of several key components: a narrow-linewidth laser, an Acousto-Optic Modulator (AOM), an Erbium-Doped Fiber Amplifier (EDFA), an optical circulator, a sensing fiber, a Balanced Photo Detector (BPD), and a Data Acquisition (DAQ) card.

The signal generation and transmission process proceed as follows: The narrow-linewidth laser generates a highly coherent continuous-wave signal. This beam is modulated into a sequence of optical pulses by the Acousto-Optic Modulator (AOM). The optical pulses are then amplified by the Erbium-Doped Fiber Amplifier (EDFA) to increase signal intensity and injected into the single-mode sensing fiber via the optical circulator. As the pulses travel through the fiber, Rayleigh backscattering, caused by microscopic inhomogeneities in the fiber’s refractive index, carries information about the fiber’s phys-

ical state. The backscattered light returns to the input end, where it is received by the Balanced Photo Detector (BPD), which converts it into analog electrical signals. Finally, the Data Acquisition (DAQ) card performs high-speed sampling and digitization of the analog signals for further processing.

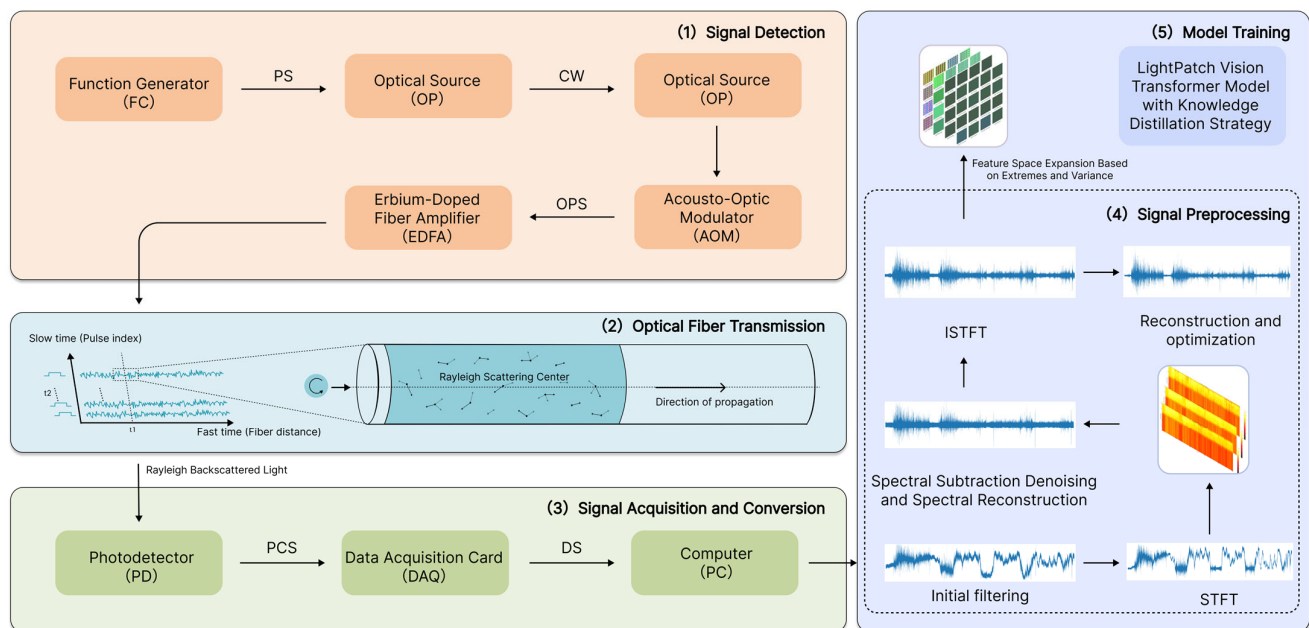


Figure 1. Workflow of the optical fiber sensing system.

The acquired data are analyzed in two temporal domains: “fast time,” which corresponds to the spatial distribution of a single pulse along the fiber, and “slow time,” which represents the temporal evolution of the signal between consecutive pulses. With high spatial resolution and strong immunity to electromagnetic interference, this system is widely applicable in areas such as perimeter security and oil and gas pipeline monitoring. However, the signals acquired by the Φ -OTDR system are typically non-stationary, bursty, and weak. Temporally, the signals are often obscured by various noise sources, resulting in a low signal-to-noise ratio (SNR) and time-varying spectral features. Spatially, effective vibration events are characterized by strong sparsity and localized abrupt changes.

The complex signal characteristics arise from multi-level interference coupling effects. At the optical level, phase noise and intensity fluctuations are caused by laser frequency instability, modulation-induced artifacts, and fiber nonlinearities. At the electronic level, thermal noise in the Balanced Photo Detector (BPD), quantization errors in the data acquisition system, and clock jitter contribute to baseline drift and sampling distortion. At the environmental level, mechanical vibrations from wind, rainfall, and geological activity generate interference signals that closely resemble real events. This complex noise environment significantly reduces phase sensitivity and obscures weak vibration signals. Addressing these challenges requires a multi-stage processing pipeline that incorporates time–frequency analysis and feature enhancement to ensure reliable event detection.

2.2. Signal Preprocessing and Image Encoding

To address the mixed characteristics of non-stationary weak signals and strong background noise in Φ -OTDR backscattering, this study proposes a multi-stage signal preprocessing architecture.

2.2.1. Signal Enhancement and Denoising

The signal enhancement and denoising process is shown in Figure 2, with detailed steps as follows:

- Initial Filtering-Power Spectral Density (PSD) analysis** using the Welch method reveals that the main energy of typical vibration events is concentrated in the 5–20 Hz frequency band. Based on this, the raw signal is processed with an equiripple FIR band-pass filter. The filter parameters are configured as follows: passband frequency of 20 Hz, stopband frequency of 5 Hz, passband ripple of 0.0575, and stopband attenuation of 0.001. This step preserves the event frequency components while suppressing low-frequency baseline drift.
- Time–Frequency analysis:** A short-time Fourier transform (STFT) with a Hamming window is applied to the time-domain signal to obtain an energy-evolution representation of events on the time–frequency plane. Direct intrusion typically corresponds to non-stationary vibrations induced by sustained mechanical actions; in the dominant band of 5–20 Hz, it manifests as a multi-second energy enhancement accompanied by a slowly evolving spectral shape over time. To achieve a more stable spectral estimate within this band and to reduce the perturbation of single-frame spectra by random noise, we set the window length to $N_{\text{win}} = 2048$ (equivalent to 4.096 s at a sampling rate of $F_s = 500$ Hz) and the hop size to $N_{\text{hop}} = 1024$ (2.048 s), i.e., a 50% overlap. A longer window improves the frequency resolution in the low-frequency range (5–20 Hz), yielding $\Delta f = F_s / N_{\text{win}} \approx 0.244$ Hz, and better characterizes slowly varying energy clusters. The 50% overlap mitigates boundary truncation effects and enhances the continuity of time–frequency energy patches across adjacent frames, thereby reducing the risk of feature attenuation and missed detections when event onsets or offsets fall near frame boundaries. The STFT is defined as follows:

$$X(t, f) = \int_{-\infty}^{\infty} x(\tau) w(t - \tau) e^{-j2\pi f \tau} d\tau \quad (1)$$

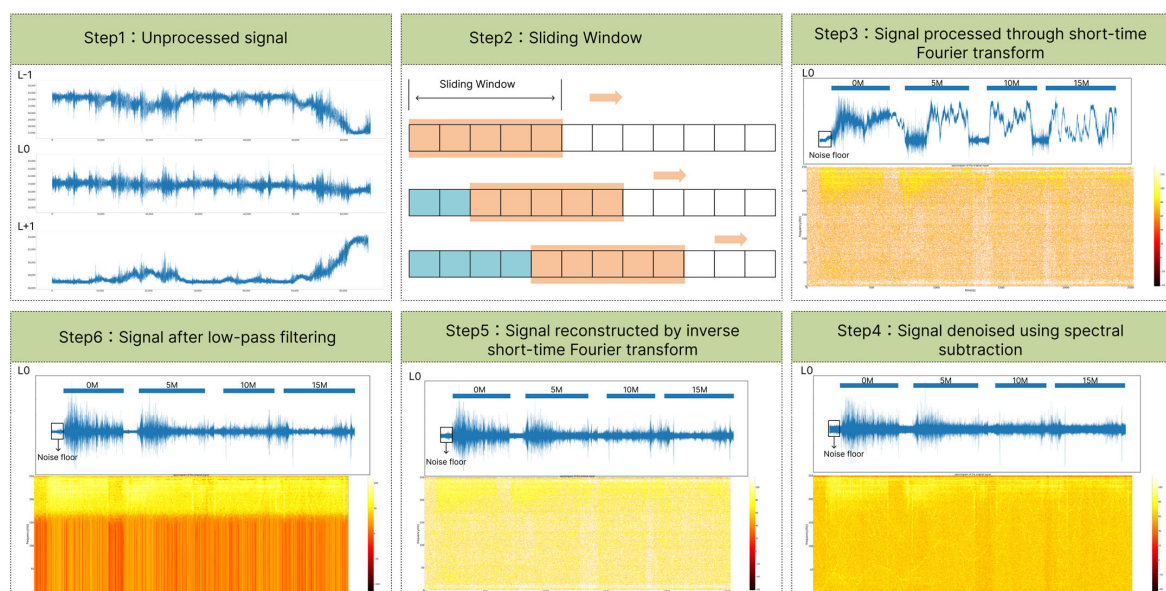


Figure 2. Flowchart of optical fiber sensing signal processing. The orange shaded region indicates the current windowed segment extracted from the signal, the blue blocks denote the previously traversed samples, and the right-pointing arrows indicate the forward shift direction of the sliding window (each row shows one shift).

Here, $x(\tau)$ is the original signal, $w(t - \tau)$ is the window function, f represents the frequency, and t denotes time.

- (c) **Spectral Subtraction Denoising:** To improve the signal-to-noise ratio of the time–frequency representation while preserving the spectral-shape structure of disturbance events, we suppress background noise in the STFT domain using spectral subtraction. The noise spectrum $H(f)$ is estimated from an initial disturbance-free baseline segment at the beginning of the recording and is subtracted frame-by-frame from the STFT of each segment. To avoid over-suppression caused by spectral subtraction, a dynamic gain-floor mechanism is introduced, formulated as follows:

$$X_{\text{clean}}(t, f) = \max(X_{\text{noisy}}(t, f) - H(f), \beta(t)X_{\text{noisy}}(t, f)) \quad (2)$$

Here, $X_{\text{noisy}}(t, f)$ denotes the noisy spectrum of the (t) -th frame, $X_{\text{clean}}(t, f)$ denotes the denoised spectrum after spectral subtraction, and $H(f)$ is the estimated noise spectrum. In this study, $H(f)$ is obtained by averaging the magnitude spectra of the first 40 consecutive STFT frames at the beginning of each recording. With $F_s = 500$ Hz and $N_{\text{hop}} = 1024$, this averaging window corresponds to approximately $40 \times 1024/500 \approx 81.9$ s. Averaging over multiple frames reduces the variance of random noise and yields a more stable estimate of $H(f)$, thereby mitigating frame-to-frame inconsistencies introduced by fluctuations in noise estimation during spectral subtraction.

Moreover, the data acquisition procedure typically includes a short baseline stage (i.e., no excavation activity and no external disturbances) before each recording starts. Therefore, the initial segment can be reasonably treated as a background-noise sample for estimating $H(f)$. In our dataset, each recording begins with a sufficiently long silent baseline segment that covers the above time scale, ensuring the stability and reproducibility of the $H(f)$ estimate. This parameter primarily determines the temporal scale of the background-estimation window; provided that the initial segment remains in a baseline state, modest variations around this scale generally do not materially change the overall statistical characteristics of $H(f)$. For implementation consistency and reproducibility, we fix 40 frames as the default setting throughout this work.

The dynamic floor coefficient $\beta(t) \in [0.1, 0.3]$ is switched in a piecewise manner according to the frame-level SNR: when the estimated SNR is low, a larger $\beta(t)$ is used to raise the minimum retained energy; when the SNR is high, a smaller $\beta(t)$ is adopted to reduce the impact on the informative spectral shape. The frame-level SNR is estimated as the ratio between the spectral energy of the current frame and that of the noise spectrum, and this estimate is then used to determine the corresponding $\beta(t)$ for each frame.

- (d) **Reconstruction and Optimization:** The signal is smoothed and reconstructed back into the original time-domain signal using the inverse short-time Fourier transform (ISTFT):

$$x(t) = \int_{-\infty}^{\infty} X(t, f) e^{j2\pi ft} df \quad (3)$$

To suppress residual baseline drift and high-frequency electronic noise, the reconstructed signal is further refined using a two-stage filtering scheme. First, a high-pass filter is applied to remove ultra-low-frequency drift; the cutoff frequency is set to $f_{\text{hp}} = 0.5$ Hz to attenuate slow baseline variations while minimally affecting the event-dominant band (5–20 Hz). Next, an n -th order Butterworth low-pass filter is employed to suppress high-frequency electronic noise and ISTFT reconstruction artifacts. The Butterworth filter exhibits a monotonic, ripple-free magnitude response in the passband, enabling a smooth roll-off while preserving the shape of the event-dominant band. This property helps reduce the influence of amplitude fluctuations and ringing on the time-domain waveform and on the

consistency of subsequent texture encoding. The magnitude response of the low-pass filter is provided in Equation (4):

$$H(f) = \frac{1}{\sqrt{1 + \left(\frac{f}{f_c}\right)^{2n}}} \quad (4)$$

Here, f_c denotes the cutoff frequency, n denotes the filter order, and f denotes frequency. We set n to 8 to balance roll-off steepness and time domain stability. The filter is applied in a zero-phase forward-backward manner to avoid phase delays that could distort event waveforms. We set the low-pass cutoff to f_c equals 75 Hz, which corresponds to a normalized cutoff of 0.3 under F_s equals 500 Hz. This choice suppresses high-frequency electronic noise and reconstruction artifacts while avoiding attenuation of informative components in the event dominant band from 5 Hz to 20 Hz.

This configurable filtering architecture significantly improves signal smoothness and distinguishability while providing the flexibility to adjust bandwidths for different vibration sources, ensuring robust detection in complex noise environments. The effectiveness of this workflow is demonstrated in Figure 3, where a comparison of the waveforms before and after processing shows substantial noise suppression and clearly distinguishable weak event features.

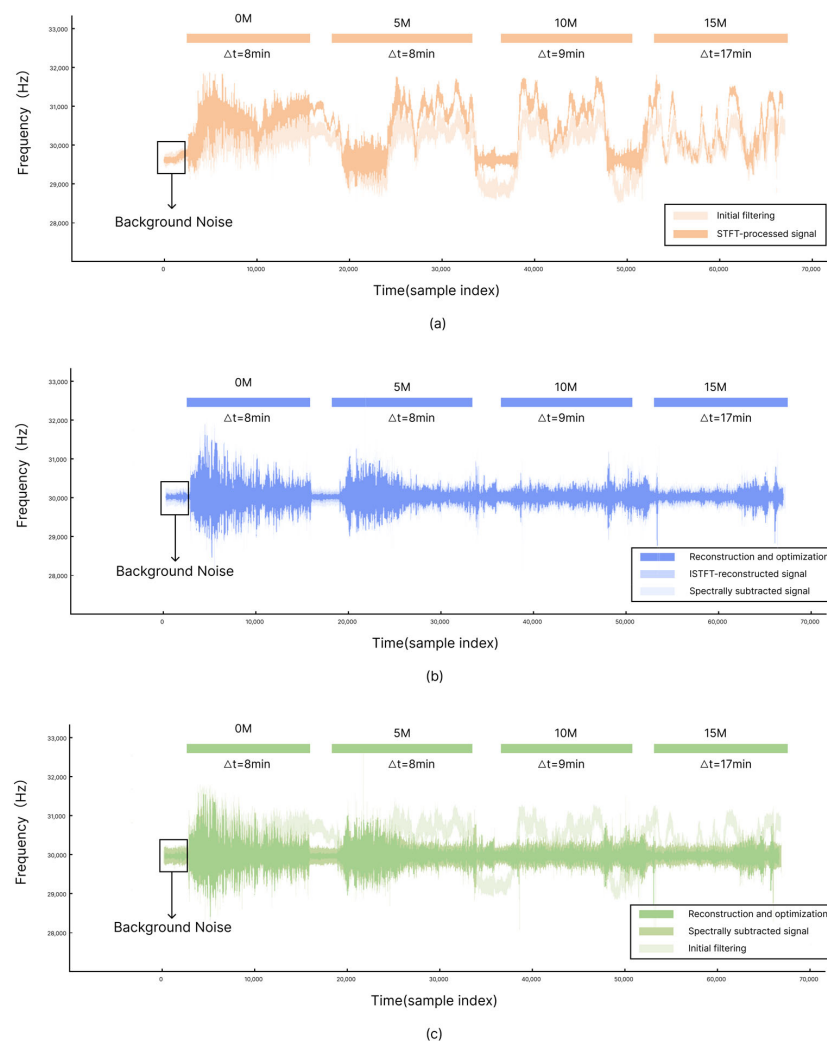


Figure 3. Illustration of multi-stage signal preprocessing effects. (a) Comparison between the initially filtered signal and the STFT-processed representation, illustrating how the STFT processing suppresses

background noise while preserving event-related components in the dominant band; (b) Comparison among the spectrally subtracted signal, the ISTFT-reconstructed signal, and the final reconstruction-and-optimization output, showing the progressive reduction of the noise floor and enhancement of event structures through denoising and reconstruction; (c) Overall comparison of the initially filtered signal, the spectrally subtracted signal, and the final reconstruction-and-optimization output, highlighting the net improvement of the full preprocessing pipeline in background-noise suppression and event separability.

2.2.2. Three-Channel RGB Image Encoding

The workflow for encoding the three-channel RGB image is shown in Figure 4.

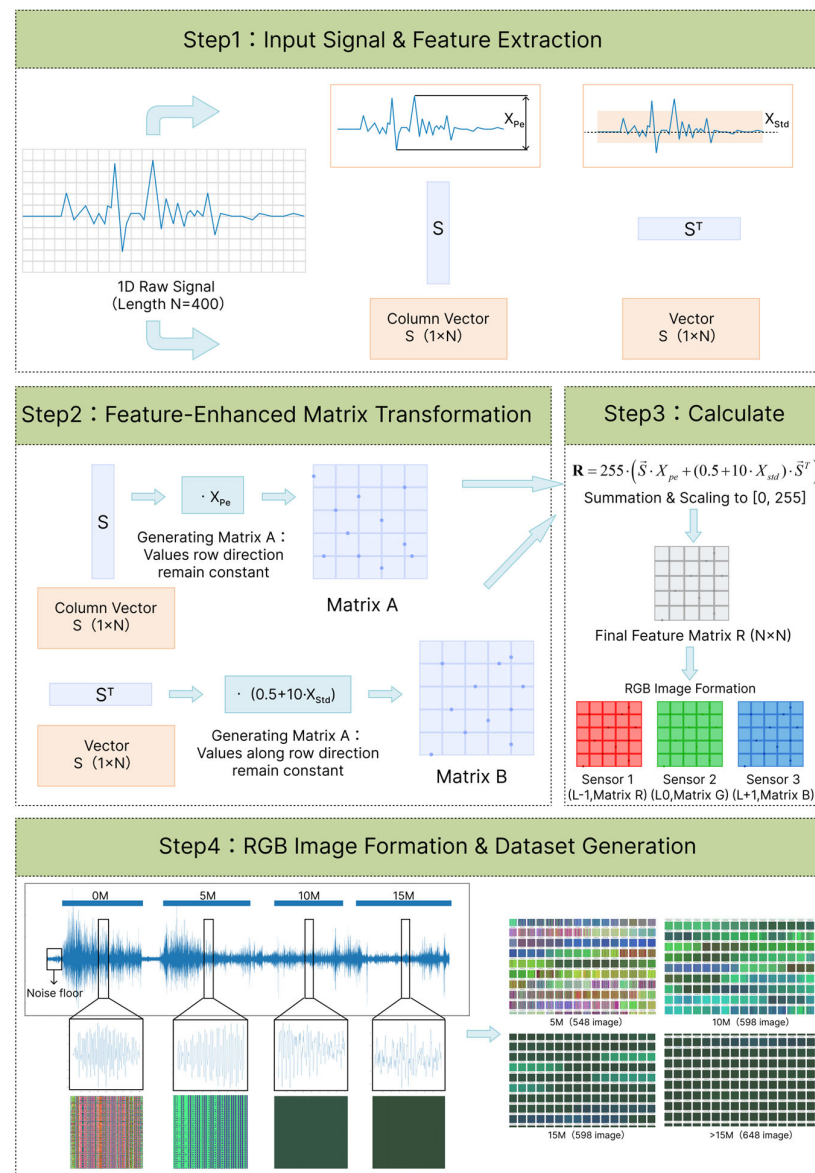


Figure 4. Schematic of the three-channel RGB image encoding process.

To provide inputs for the distance-classification model, the one-dimensional vibration sequence is transformed into a three-channel RGB image. This design reflects the physical propagation characteristics of vibration waves in soil, where energy decays with distance. Signals from three adjacent sensing points ($L - 1$, L_0 , $L + 1$) are assigned to the R, G, and B channels, respectively. This encoding allows the resulting image to represent the spatial attenuation pattern of the vibration.

For each sensor point with a signal length of N ($N = 400$), the one-dimensional signal x is transformed into an $N \times N$ two-dimensional feature matrix R , where the transformation formula is defined as follows:

$$R = 255 \left(\vec{S} \cdot X_{pe} + (0.5 + 10 \cdot X_{std}) \cdot \vec{S}^T \right) \quad (5)$$

Here, $\vec{S}$ denotes the column-vector form of the normalized signal, and $\vec{S}^T$ is its transposed row-vector form. The matrix M is constructed by superposing a column-wise structural term $\vec{S} \cdot X_{pe}$ and a row-wise structural term $(0.5 + 10X_{std}) \cdot \vec{S}^T$, which expands the temporal profile of $\vec{S}$ along the column and row directions, respectively, to form a two-dimensional structural texture. Here, X_{pe} is the peak-to-peak value of the segment, and X_{std} is the standard deviation; both are computed from the original (pre-normalization) segment S to preserve information on the absolute intensity and fluctuation scale of the event, whereas $\vec{S}$ is used solely to keep the numerical range stable during matrix construction. As global statistical features, X_{pe} and X_{std} perform intensity weighting and volatility modulation, respectively, by applying multiplicative and bias adjustments to the above structural terms to enhance discriminative characteristics. The constant 255 is used to scale the values of matrix M to the standard 8-bit pixel intensity range $[0, 255]$. To ensure numerical stability and reproducibility, the time-domain signal $\vec{S}$ is first Min–Max normalized to $[0, 1]$ before constructing M . During pixelization, M is then saturated clipped to $[0, 255]$. Because the additive terms include weighted/bias modulation by X_{pe} and X_{std} , the values of M may exceed $[0, 1]$; therefore, saturation clipping is applied at the pixelization stage to guarantee a stable 8-bit mapping.

Specifically, $\vec{S} \cdot X_{pe}$ produces a matrix that is constant along the row direction and is scaled by X_{pe} , thereby retaining absolute intensity information. In contrast, $(0.5 + 10X_{std}) \cdot \vec{S}^T$ generates a matrix that is constant along the column direction and is modulated by a volatility-related factor to emphasize non-stationary variations; the constant 0.5 ensures a non-negligible baseline contribution when X_{std} is small, while the coefficient 10 amplifies the influence of volatility. The resulting two-dimensional texture differences are primarily governed by the temporal structure of $\vec{S}$ —such as the timing and density of impulsive components, intermittency, and the evolution of the energy envelope—which, after row/column expansion, correspond to distinguishable stripe- and block-like geometric patterns. Meanwhile, X_{pe} and X_{std} mainly serve as intensity and volatility modulators to improve robustness.

By stacking signals from adjacent sensing points into three channels, the spatial attenuation gradient is further transformed into cross-channel contrastive structures. As a result, differences in temporal patterns and spatial attenuation associated with different intrusion mechanisms can be reflected in separable texture representations, rather than being captured merely as amplitude variations.

Finally, the enhanced matrices generated from the three sensing points are stacked into the R, G, and B channels, forming an $N \times N \times 3$ RGB image. In this study, the time-domain length of each sample segment is set to $N = 400$, resulting in a per-sample image size of $400 \times 400 \times 3$, and a total of 2392 images are constructed. To match the network input size, all images are uniformly resampled to a fixed resolution (224×224) before training and inference, using bilinear interpolation without cropping. This operation primarily changes the discrete sampling density of the textures while preserving the one-to-one correspondence between the three channels and adjacent sensing points. Its potential impact on distance classification is further discussed in Section 4.

2.3. Methodology for Network Model

2.3.1. LightPatch Vision Transformer (LP-ViT)

While CNNs enable end-to-end feature extraction, they have fundamental limitations in processing Φ -OTDR fringe patterns. Specifically, restricted receptive fields fail to capture long-range dependencies (particularly in the 5–15 m range), and intermediate dimensional compression often discards weak disturbance signals. To overcome these challenges, this study proposes a hybrid architecture, the LP-ViT, as illustrated in Figure 5.

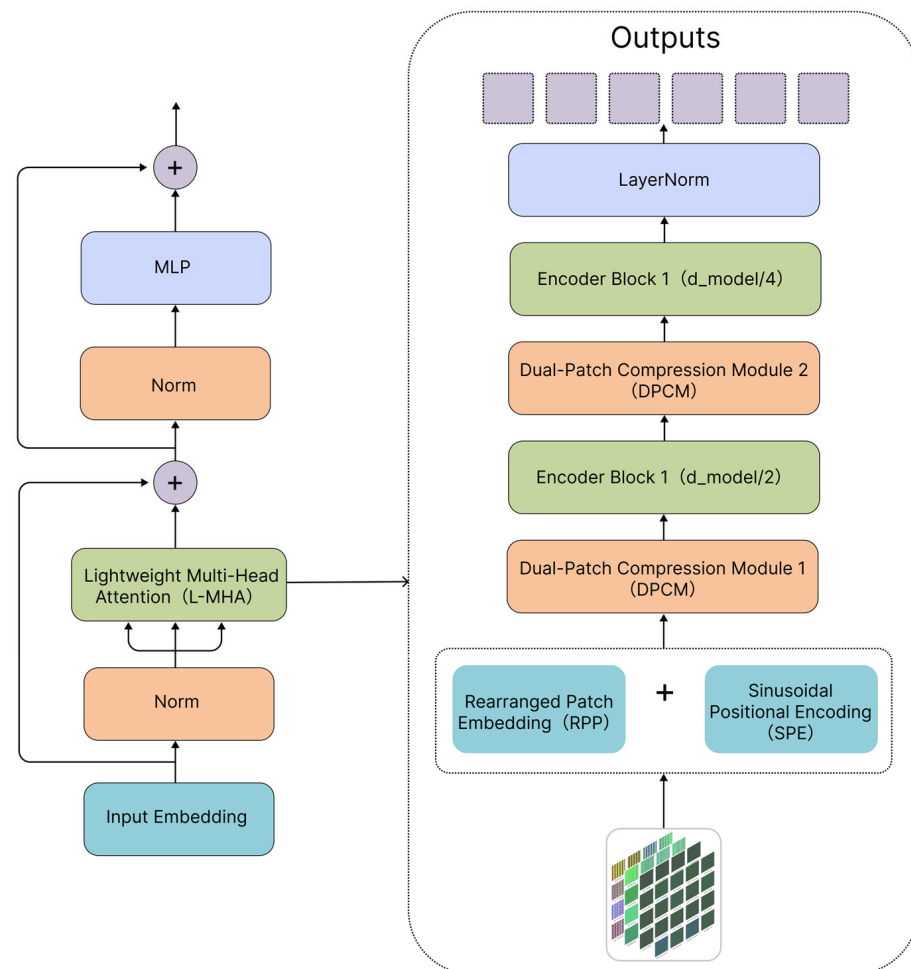


Figure 5. LightPatch Vision Transformer network architecture.

The architecture features a hierarchical encoding design utilizing a Dual-Path Compression Module (DPCM) to progressively reduce feature dimensionality to $d_{\text{model}}/2$ and $d_{\text{model}}/4$, while ensuring stability via LayerNorm. The core Lightweight Multi-Head Attention (L-MHA) module integrates Sinusoidal Positional Encoding (SPE) to retain global spatial structure and Relative Patch Positioning (RPP) to enhance local feature extraction. Furthermore, a dual-path feature interaction mechanism at the input embedding layer incorporates CNN-style local perception. By establishing arbitrary pixel-wise dependencies through self-attention, this architecture effectively models long-range correlations and fuses multi-scale features from fine to coarse granularity.

2.3.2. Knowledge Distillation for Lightweight Deployment

To bridge the gap between high-performance modeling and edge deployment constraints in Φ -OTDR monitoring, this study employs a knowledge distillation optimization scheme, as shown in Figure 6. In this teacher–student framework, the original LP-ViT acts

as the teacher, while a pruned and channel-sparsified version serves as the student. The core innovation lies in a dual knowledge transfer mechanism that captures both decision boundaries and internal feature representations.

- (a) **Soft Label Knowledge Transfer:** A temperature-scaled Softmax function is used to enable the smooth transfer of prediction distributions. By introducing a temperature parameter T , soft labels preserve the inter-class relational information learned by the teacher model, which is crucial for distinguishing similar vibration patterns in Φ -OTDR data. The temperature-scaled Softmax function is defined as follows:

$$\text{Softmax}(z_i, T) = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \quad (6)$$

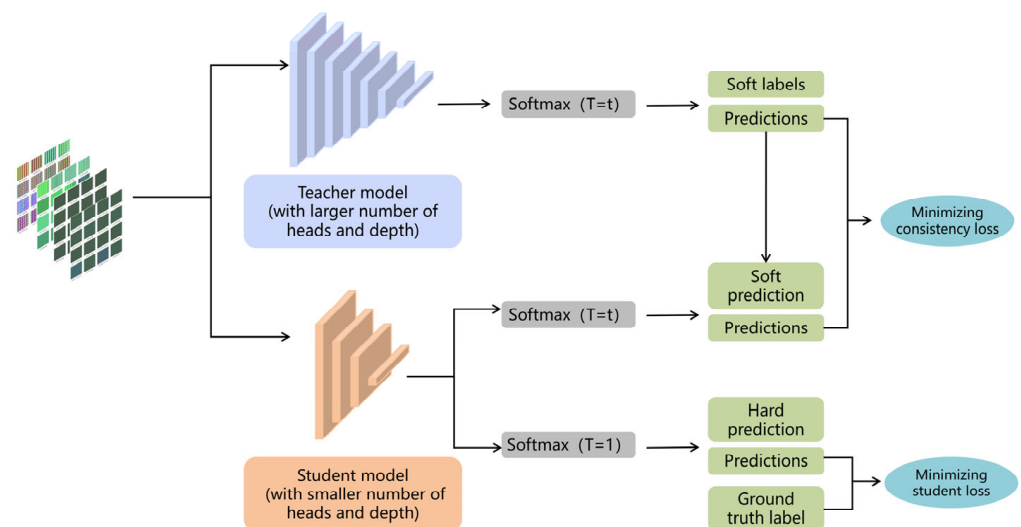


Figure 6. Framework for ViT model optimization using a teacher–student approach and knowledge distillation.

Here z_i denotes the logit for class i , and T is the temperature parameter. As T increases, the output probability distribution becomes smoother, thereby facilitating stable knowledge transfer.

- (b) **Intermediate-Layer Feature Alignment:** To enhance representational capacity, an intermediate-layer feature alignment strategy is introduced. This mechanism imposes constraints on both attention matrices (A) and token representations (T), allowing the student to mimic the teacher’s internal dynamics and capture global dependencies. The specific loss function is defined by the following:

$$L_{align} = \lambda_1 \|A_{teacher} - A_{student}\|_F^2 + \lambda_2 \|T_{teacher} - T_{student}\|_2^2 \quad (7)$$

Here, $A_{teacher}$ and $A_{student}$ represent the attention matrices of the teacher and student models at a given layer, while $T_{teacher}$ and $T_{student}$ denote their respective token representations. $\|\cdot\|_F$ denotes the Frobenius norm, which is equivalent to the L2 norm. The hyperparameters λ_1 and λ_2 control the balance between the two loss components.

This staged approach ensures that the student model inherits both decision boundaries and fine-grained feature representations, maximizing recognition performance while achieving model compression.

3. Result

3.1. Dataset and Experimental Setup

In this study, we conducted field experiments using a distributed fiber-optic sensing system. The sensing fiber was installed near the pipeline and mechanically coupled to the ground surface. The system provided 8192 monitoring points and sampled at 500 Hz, with the spatial layout shown in Figure 7.

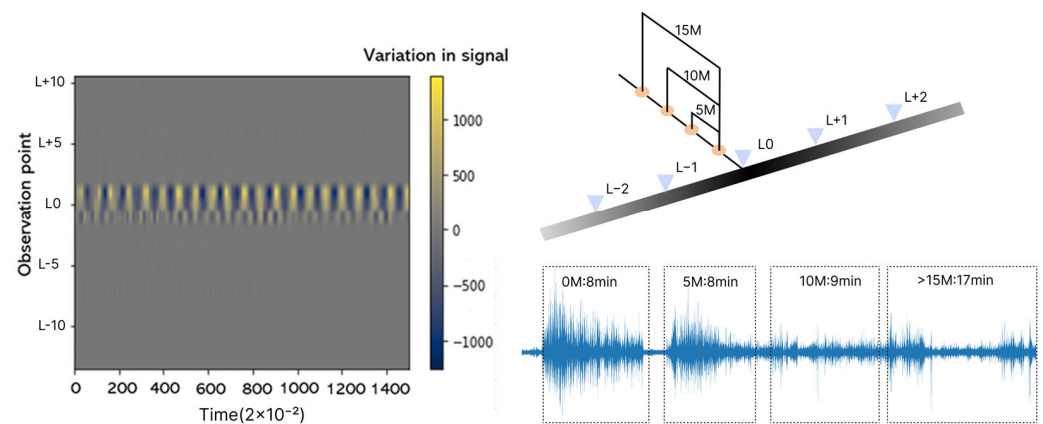


Figure 7. Relationship between pipeline vibration signals and operation distance. L0 denotes the reference observation point. L plus k and L minus k indicate the k-th location relative to L0 along the fiber, with the minus sign representing the opposite direction. In the left heatmap, color encodes the magnitude of signal variation as indicated by the colorbar. In the schematic, the markers denote the measurement locations associated with the 0 M, 5 M, 10 M, and 15 M distance conditions, and the dashed boxes highlight the corresponding waveform segments.

To capture intrusion-induced vibrations at different stand-off distances, excavation was performed at approximately 0 m, 5 m, 10 m, and 15 m from the sensing fiber. Data were recorded sequentially as a single continuous time series. The corresponding time intervals were 0 to 8 min for 0 m, 9 to 17 min for 5 m, 18 to 27 min for 10 m, and 28 to 45 min for 15 m. Before the continuous recording began, an approximately 2 min silent baseline was collected to stabilize the system and the environment and to characterize background noise. An approximately 1 min silent interval was also inserted between adjacent distance conditions to allow condition switching and stabilization. These silent intervals were used only as a reference and were excluded from sample generation and label assignment. Samples were extracted only from the annotated time intervals for each distance condition. For unified downstream processing, the condition-specific segments were concatenated in chronological order to form a continuous signal sequence.

To construct model inputs, we used the annotated event-affected region and selected three adjacent monitoring points near this region as the spatial sampling channels, namely L minus 1, L0, and L plus 1. The remaining monitoring points were used for event localization and background reference and were not used for image construction. During preprocessing, the continuous signal was segmented with a sliding window. The window length N equals 400 samples, corresponding to 0.8 s, and the stride equals 200 samples, corresponding to 0.4 s. Windows were extracted only from the time intervals assigned to each distance condition. For each window, we computed a time–frequency representation using the short-time Fourier transform. We used N_{win} equals 2048, and N_{hop} equals 1024, and then mapped the three adjacent point representations to three channels to form an RGB input image. Using this procedure, we generated 2392 two-dimensional time–frequency images for training and evaluation. The dataset contains four distance conditions, namely 0 m, 5 m, 10 m, and 15 m. In the training set, the number of images is 451, 439, 481, and 497 for

the four conditions, respectively. In the test set, the corresponding numbers are 132, 129, 127, and 136.

It should be noted that the number of valid samples for each distance condition was not generated in strict proportion to the total recording duration. To limit redundancy caused by highly overlapping windows and to maintain label consistency, we prioritized segments that exhibited clearer and more stable event energy within each condition interval when constructing images. As a result, the effective sample counts across conditions do not scale linearly with recording length.

The dataset is labeled by the on-site stand-off distance between the excavation activity and the sensing fiber, and the four distance conditions are further translated into four response tiers for warning deployment. Specifically, 0 m corresponds to an ultra-close condition where the excitation source is directly above the fiber and thus has the highest response priority; 5 m and 10 m represent near-range direct intrusion and high-risk operations, respectively; and 15 m corresponds to relatively distant background construction. This distance-to-threat mapping enables the model outputs to be directly used for online warning prioritization and response decision-making.

Under a unified data splitting and evaluation pipeline, all compared models were trained using the same set of hyperparameters, and regularization strategies were applied to mitigate overfitting and improve generalization. The experimental configurations—including the selected hyperparameters, hardware specifications, and dataset statistics—are summarized in Table 1.

Table 1. Optimal configuration of model parameters.

Parameter	Value
Image Size	224×244
Initial Learning Rate	3×10^{-5}
Batch Size	32
Epoch	100
Training Framework	Pytorch
Optimizer	AdamW
Weight Decay	1×10^{-2}
Label Smoothing	0.1

3.2. Model Comparison Experiments

To evaluate the performance of various neural architectures for Φ -OTDR signal classification, this study developed a comprehensive benchmarking framework that includes mainstream CNNs (AlexNet, VGG, and ResNet) and Transformer variants (ViT, Swin, and DeiT). Given the inherent long-range spatial dependencies of Φ -OTDR vibration signals, the study focuses on comparing the efficacy of self-attention mechanisms with traditional local convolutions.

To avoid information leakage from temporally adjacent windows, we build the training and test sets using a time-blocked split. Under this fixed split, we repeat training with multiple random seeds to assess result stability, and the sample counts and class distributions are provided in Section 3.1. We report overall accuracy and macro-F1 as the primary evaluation metrics. To further reflect recognition performance at different distances, Table 2 also lists the per-class correct recognition rates for 0 M, 5 M, 10 M, and 15 M, defined as the proportion of correctly predicted samples within each class. All metrics are summarized in Table 2 using mean and standard deviation across random seeds.

Table 2. Performance comparison of different neural network architectures in Φ -OTDR signal classification tasks.

Model	Architecture	0 M Mean Accuracy	5 M Mean Accuracy	10 M Mean Accuracy	15 M Mean Accuracy	Macro Mean Accuracy	Accuracy Mean $\pm$ std	Macro-F1 Mean $\pm$ std	Macro-F1 95% Confidence Interval	Folds $\times$ Seeds
ViT	Transformer	0.866	0.908	0.835	0.920	0.885	0.908 $\pm$ 0.009	0.891 $\pm$ 0.009	[0.8907, 0.9030]	5 $\times$ 5
VGG13	CNN	0.933	0.907	0.680	0.912	0.858	0.865 $\pm$ 0.007	0.857 $\pm$ 0.008	[0.8473, 0.8673]	5 $\times$ 5
VGG16	CNN	0.839	0.893	0.694	0.932	0.839	0.860 $\pm$ 0.008	0.849 $\pm$ 0.010	[0.836, 0.862]	5 $\times$ 5
VGG11	CNN	0.789	0.892	0.662	0.931	0.819	0.844 $\pm$ 0.015	0.832 $\pm$ 0.019	[0.8083, 0.8558]	5 $\times$ 5
DeiT-Tiny	Transformer	0.7186	0.8424	0.7329	0.8853	0.7948	0.818 $\pm$ 0.012	0.808 $\pm$ 0.015	[0.7889, 0.8277]	5 $\times$ 5
Swin_Tiny	Transformer	0.560	0.871	0.719	0.877	0.757	0.793 $\pm$ 0.013	0.773 $\pm$ 0.019	[0.7493, 0.7984]	5 $\times$ 5
ResNet18	CNN	0.261	0.962	0.874	0.919	0.754	0.815 $\pm$ 0.010	0.752 $\pm$ 0.019	[0.7284, 0.7764]	5 $\times$ 5
ResNet34	CNN	0.668	0.764	0.659	0.784	0.719	0.735 $\pm$ 0.010	0.723 $\pm$ 0.013	[0.7067, 0.7400]	5 $\times$ 5
ResNet50	CNN	0.120	0.787	0.554	0.946	0.602	0.703 $\pm$ 0.013	0.598 $\pm$ 0.021	[0.5716, 0.6255]	5 $\times$ 5

Note: Bold values indicate the best-performing model (highest score) for each metric.

Under the unified five-fold cross-validation protocol with five repeated runs per fold, the results show that transformer-based models generally outperform most CNN baselines. This suggests stronger global modeling capability and more stable overall performance. In particular, the standard ViT achieves the best and most consistent results, with a mean accuracy of 0.9084 ± 0.0091 , and it also leads on the overall metrics. By comparison, the VGG family is the strongest among the CNN baselines, but it remains inferior to the standard ViT overall. The other CNN baselines perform weaker on this task. These findings support using the standard ViT as the reference baseline for subsequent lightweight design and knowledge distillation optimization.

To assess whether the observed performance differences among models are statistically robust under the unified evaluation protocol, we perform a nonparametric significance analysis on macro-F1 using paired results from repeated cross-validation with multiple random initializations. Specifically, we use five-fold cross-validation, and within each fold, we repeat training and evaluation with five independent random seeds, yielding 25 paired runs per model under the same data partitions. We first apply the Friedman test for an overall comparison across models, and then conduct pairwise Wilcoxon signed rank post hoc tests with Holm correction for multiple comparisons. The results show that the statistical ranking agrees with the mean macro-F1 reported in the main table. ViT achieves the best macro-F1, and its performance differences relative to each competing model remain statistically significant after multiple comparison correction. Detailed settings and a summary of the test results are reported in Table 3.

Table 3. Statistical significance analysis of model macro-F1.

Optimal Model	Comparison Model	Optimal Mean	Compare Mean	Mean Difference
ViT	DeiT-Tiny	0.9019	0.8083	0.0936
ViT	ResNet18	0.9019	0.7524	0.1494
ViT	ResNet34	0.9019	0.7234	0.1785
ViT	ResNet50	0.9019	0.5986	0.3033
ViT	Swin_Tiny	0.9019	0.7739	0.1280
ViT	VGG11	0.9019	0.8320	0.0698
ViT	VGG13	0.9019	0.8573	0.0446
ViT	VGG16	0.9019	0.8496	0.0523

To better understand the sources of performance differences across architectures, we aggregate the confusion matrices of all models under five-fold cross validation with five repeated runs per fold, as shown in Figure 8. Overall, most misclassifications occur between adjacent distance tiers. In particular, 10 M to 15 M and 10 M to 5 M repeatedly appear as high-frequency confusions across multiple models. Some CNN baselines show a stronger cross-tier bias. For example, ResNet18 and ResNet50 more often misclassify 0 M as 15 M,

which substantially degrades their macro-averaged performance. In contrast, the standard ViT still mainly confuses adjacent tiers. Its errors are more dispersed and less severe overall, which leads to better overall metrics and higher stability.

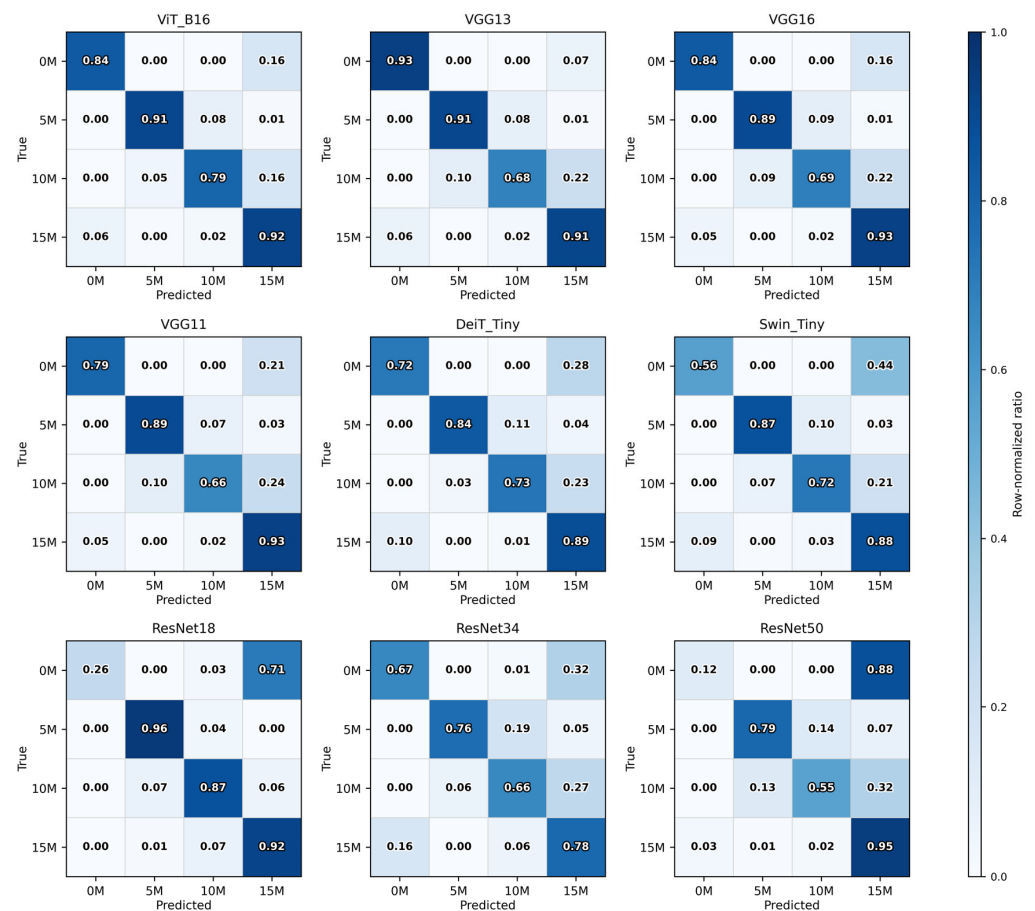


Figure 8. Comparison of confusion matrices across network architectures. (a) Row-normalized aggregated confusion matrix of ViT; (b) Row-normalized aggregated confusion matrix of VGG13; (c) Row-normalized aggregated confusion matrix of VGG16; (d) Row-normalized aggregated confusion matrix of VGG11; (e) Row-normalized aggregated confusion matrix of DeiT-Tiny; (f) Row-normalized aggregated confusion matrix of Swin-Tiny; (g) Row-normalized aggregated confusion matrix of ResNet18; (h) Row-normalized aggregated confusion matrix of ResNet34; (i) Row-normalized aggregated confusion matrix of ResNet50.

Based on the above comparisons, we select the standard ViT as the backbone for subsequent architectural modifications and lightweight design. This choice preserves a strong performance upper bound while providing a higher-performing starting point for model compression and distillation toward edge deployment.

3.3. Comprehensive Evaluation of Recognition Performance and Edge Deployability

To systematically evaluate the proposed method in terms of the overall trade-off between recognition performance and edge deployability, this section presents three groups of comparisons under a unified data split, input specification, and training setting:

- Performance–complexity trade-off analysis: The computational complexity of the standard ViT, with 85.65 M parameters, severely limits its deployability on edge nodes. To address this, the proposed LP-ViT architecture optimizes the trade-off between accuracy and efficiency. As shown in Table 4, under a fixed train/validation split (blocked split), we repeat training with multiple random seeds and statistically

summarize the results. By incorporating structured pruning and lightweight attention mechanisms, LP-ViT reduces the parameter count from 85.65 M to 2.88 M, a 96.6% reduction, and lowers VRAM usage from 1368.71 M to 95.6 M. To achieve further lightweighting, knowledge distillation was applied to create a student model. Results show that this model reduces parameters to 0.37 M and VRAM usage to 56.88 MB, while decreasing inference latency to 5.5 ms. With a sustained accuracy of 87.96%, the student model strikes the optimal balance between computational efficiency and sensing reliability for real-time edge applications.

- (b) Quantitative validation of the contribution of knowledge distillation: To validate the crucial role of knowledge distillation in addressing the capacity limitations of lightweight models, this study compared the performance of the student model under two paradigms: “Training from Scratch” and “Teacher-Guided,” using identical architectural configurations. As shown in Table 5, relying solely on the student network’s limited parameter capacity hampers its ability to capture subtle features in fiber optic vibration signals, resulting in lower classification accuracy. However, by integrating the distillation mechanism, which benefits from the inter-class correlation information in the “soft targets” provided by the teacher, the student model’s feature extraction capability is significantly enhanced. As a result, test accuracy improves to 87.96%, demonstrating that the distillation strategy effectively facilitates high-fidelity knowledge transfer from the complex teacher to the compact student network.
- (c) Edge-side deployment efficiency comparison with existing lightweight Φ -OTDR models: To quantify the online deployability of different models under a unified edge-evaluation benchmark, we note that existing edge-oriented lightweight network designs for Φ -OTDR applications are still predominantly CNN-based. Therefore, we select three representative recent lightweight CNN models (STFT-AECNN, DVS-CNN-4, and DWConv-3) as the primary baselines, ensuring engineering reproducibility and like-for-like comparability. All models are evaluated on the same hardware under identical data preprocessing and input-resolution settings. A consistent set of metrics—including inference latency, throughput, peak GPU memory usage, and model size—is used for statistical reporting. The results are summarized in Table 6.

Table 4. Comparison of computational efficiency and classification performance between standard ViT and proposed LP-ViT models.

Indicator	Standard ViT	LP-ViT (Teacher Model)	Student Model
Total parameter amount (MB)	85.65	2.88	0.37
Accuracy of training at the 100th epoch	90.58%	89.9%	89.06%
Accuracy of the 100th Epoch test	89.83%	88.65%	87.96%
Training time per epoch (ms)	42	11.5	5.5
VRAM allocation (MB)	1368.71	95.6	56.88
VRAM reserved (MB)	8522	402.7	219.6

Table 5. Impact of knowledge distillation on student model performance.

Model	Parameters (MB)	Test Set Accuracy (%)	Precision (%)	Recall (%)
Teacher Model	2.88	88.65	89.38	88.76
Baseline Student	0.37	81.58	82.14	80.83
Final Student Model	0.37	87.96	88.23	87.51

Table 6. Edge-deployment performance comparison with recent lightweight models for Φ -OTDR.

Model	Latency Mean (ms)	Latency p95 (ms)	Throughput (sps)	Peak VRAM Reserved (MB)	Best Acc
Student	3.3110	4.1469	263.36	506	0.8906
STFT-AECNN	3.1802	4.4267	127.53	786	0.7937
DVS-CNN-4	1.2227	1.9474	175.59	360	0.7725
DWConv-3	1.1061	2.2557	222.36	316	0.6896

From a deployment-efficiency perspective, DVS-CNN-4 and DWConv-3 achieve lower mean inference latency (1.22 ms/sample and 1.11 ms/sample, respectively) and smaller model sizes (0.0249 M and 0.0015 M parameters), but their best test accuracies are only 0.7725 and 0.6896, respectively. By contrast, while maintaining a best test accuracy of approximately 0.89, our student model still delivers edge-side inference efficiency with a mean latency of 3.31 ms·sample^{−1}, a p95 latency of 4.15 ms·sample^{−1}, and a throughput of 263.36 samples·s^{−1}, while keeping the model size at 0.9251 M parameters with a 3.55 MB checkpoint. Therefore, the student model is not designed for extreme lightweighting; instead, it preserves recognition reliability while maintaining acceptable real-time inference performance and storage/memory overhead, making it more suitable for continuous online monitoring on resource-constrained edge nodes.

3.4. Engineering Rationale and Robustness Validation of Distance-Based Grading Thresholds

Because Φ OTDR offers high sensitivity and long-range continuous monitoring, vibration measurements along the fiber often include both potential intrusion events and background disturbances from multiple sources. Disturbances occurring farther from the pipeline are more likely to be dominated by activities that are not directly related to the pipeline. In contrast, near-range disturbances are more likely to originate from operations close to the pipeline or from external forces acting on the pipeline itself. Therefore, near-range events should be assigned higher priority for response scheduling.

Distance-based zoning is also consistent with institutional constraints in pipeline corridor safety management. Regulations and standards commonly define the near pipeline corridor as a high-sensitivity control zone. They prohibit or strictly restrict excavation, piling, and other construction activities that may threaten pipeline safety within a 5 m buffer on both sides of the pipeline [25]. Accordingly, activities within 5 m are mapped to a higher response urgency level. From the risk assessment perspective of pipeline integrity management, operation, and maintenance practices require potential threat events to be identified, graded, and prioritized to support response decisions under limited resources [26]. In this study, all distance conditions were collected on the same pipeline section under a controlled field setting, so differences in consequence severity are relatively limited. Therefore, distance can serve as an actionable indicator of event relevance to the pipeline and of response urgency, which enables a deployable graded output.

To better align the graded outputs with on-site response workflows, we introduce two auxiliary boundaries at 10 m and 15 m beyond the 5 m high-sensitivity control zone. This design separates three response regimes, namely near-field strongly coupled disturbances, a transitional influence region, and far-field weakly coupled background activity. Under common third-party construction activities such as excavation, tamping, and piling, the vibration energy coupled into the pipeline typically decays markedly with distance. Near-range disturbances are therefore more likely to involve stronger coupling, higher pipeline stress, or even direct contact hazards. Accordingly, events within 5 m to 10 m are assigned a relatively high verification priority. When the distance increases from 10 m to 15 m, the observed signatures more often reflect indirectly coupled vibrations after propagation attenuation. In this range, a medium priority strategy based on continuous monitoring

plus confirmatory re-check is generally appropriate. In our controlled field setting, the 15 m condition serves as a practical approximation of a far-field weak-coupling background tier. Such events are more likely to correspond to activities with lower relevance to the pipeline and can be treated as a low priority in response planning.

The proposed distance-grading thresholds are designed to produce discrete tiered outputs that are consistent with response prioritization, and their engineering meaning mainly lies in an operational representation of “relative risk/response urgency.” To mitigate boundary sensitivity introduced by discrete thresholds and to further capture the continuous variation in threat intensity, we introduce—beyond discrete grade recognition—a continuous score S derived from the model’s probabilistic outputs to represent the expected threat intensity across three distance tiers. We subsequently validate the robustness of the grading rule via a threshold-perturbation stability analysis. Both the stability and correlation analyses are computed from prediction outputs on an independent test set, in order to avoid information leakage into the training process.

Among the four distance conditions, 0 M represents ultra close disturbances and, in engineering practice, can be assigned directly to the highest priority response queue. We therefore use a continuous score S to characterize relative threat intensity and decision boundaries among the remaining three tiers, namely 5 M, 10 M, and 15 M. This score is also used in the subsequent threshold perturbation stability analysis. Let $P(k|x)$ denote the predicted probability that a sample x belongs to distance tier $k \in \{15\text{ M}, 10\text{ M}, 5\text{ M}\}$. Since our threat-level definition follows an ordinal relation in which “closer distance implies higher response priority,” the weights must satisfy $\omega_{5\text{M}} > \omega_{10\text{M}} > \omega_{15\text{M}}$ to ensure that S increases when the model assigns higher probability to nearer-distance tiers. Accordingly, the continuous score is defined as follows:

$$S(x) = \sum_k w_k P(k|x) \quad (8)$$

Based on the prediction outputs on the independent test set, the mean S values for the 15 M, 10 M, and 5 M classes are 0.813, 1.747, and 2.787, respectively, exhibiting a stable increasing gradient. The Spearman rank correlation coefficient between S and the distance-tier labels is 0.804. To examine the dependence of this ordering on the choice of weights, we evaluated multiple weight combinations while maintaining the constraint $\omega_{5\text{M}} > \omega_{10\text{M}} > \omega_{15\text{M}}$, as reported in Table 7; the Spearman coefficient remains stable in the range 0.796–0.809. Because different weight settings amount to a monotonic re-parameterization of the same probability vectors and do not alter their relative probabilistic structure across classes, these results indicate that the ordering consistency is primarily driven by the model’s probability allocation patterns across the three classes rather than by any specific weight values.

Table 7. Stability of rank correlation for the continuous threat score under different conditions.

Weight Combination	5 M	10 M	15 M	Spearman
1, 2, 3	3	2	1	0.804
0, 1, 2	2	1	0	0.796
1, 3, 5	5	3	1	0.805
1, 5, 10	10	5	1	0.809

To make the continuous score directly applicable to grading decisions in deployment, we specify reproducible decision boundaries in the S -space and evaluate how small boundary adjustments affect the grading outcomes. Using the midpoints between the mean S values of adjacent tiers as the baseline boundaries yields a low–medium boundary of 1.280 and a medium–high boundary of 2.267. When a simultaneous offset of ± 0.10 is applied to both boundaries around the baseline, the proportions of samples whose assigned response

tiers remain unchanged are 0.952 and 0.979, respectively. These results indicate that within the range of threshold fine-tuning typically permitted in engineering deployment, the overall graded outputs remain stable, and the changes are mainly concentrated on samples with scores close to the boundaries. This provides empirical evidence for the usability and robustness of the distance-based grading outputs in response-prioritization scenarios.

3.5. Ablation Studies

To validate the impact of key design choices on recognition performance and edge deployment, we conduct ablation studies from two perspectives: recognition effectiveness and deployment efficiency. For the input representation and critical architectural components, performance is evaluated using accuracy (Acc) and macro-F1. For the distillation strategy and the capacity configuration of the student model, we assess performance using Acc together with deployment-related metrics, such as inference latency and memory footprint.

- (a) **Ablation on signal preprocessing:** With the LP-ViT architecture, data split, and training settings kept identical, we perform an ablation study on key components of the signal preprocessing pipeline. The compared variants include the following: (i) the full pipeline (FIR band-pass filtering + STFT + spectral-subtraction denoising + ISTFT reconstruction with smoothing/post-processing), (ii) removing the FIR band-pass filter, (iii) removing spectral-subtraction denoising, and (iv) removing ISTFT reconstruction and smoothing/post-processing, to evaluate how each preprocessing component affects threat-level recognition performance. The results are reported in Table 8.

Table 8. Ablation results of the signal preprocessing module.

Setting	Acc (%)	Macro-F1 (%)
Full (FIR + STFT + SpecSub + ISTFT + Smooth)	89.96	89.61
w/o FIR band-pass	87.42	85.35
w/o Spectral Subtraction	87.21	85.73
w/o ISTFT + smoothing	84.84	82.70

As shown in Table 8, the full preprocessing pipeline achieves the best performance (Acc = 89.96%, macro-F1 = 89.61%). Removing the FIR band-pass filter or spectral-subtraction denoising leads to a slight degradation, indicating that both components help suppress background interference and improve the separability of the time–frequency representation. In contrast, removing ISTFT reconstruction and smoothing/post-processing results in a much more pronounced drop, suggesting that this step is more critical for stabilizing time–frequency texture structures and mitigating the effects of residual noise and baseline fluctuations. Therefore, the full signal preprocessing pipeline is adopted as the default input-enhancement front end in subsequent experiments.

- (b) **RGB Dimension Expansion Ablation:** Under the same LP-ViT architecture, data split, and training settings, we evaluate the impact of different input representations by using (i) three-channel RGB, (ii) gray (a grayscale image replicated across three channels), and (iii) single-channel variants replicated across channels (R-only, G-only, and B-only). We then compare the recognition performance under these channel configurations for threat-level classification. The results are summarized in Table 9.

Table 9. Ablation results of the RGB representation module.

Setting	Acc (%)	Macro-F1 (%)
RGB	89.96	89.61
Gray	78.32	72.35
R-only	85.89	82.80
G-only	77.47	70.84
B-only	65.05	57.04

As shown in Table 9, the three-channel RGB input achieves the best performance (Acc = 89.96%, macro-F1 = 89.61%). When the input representation degenerates to a single channel or to grayscale replication, the recognition performance decreases noticeably to varying degrees. This indicates that multi-channel fusion provides complementary information and enhances class separability. Among the single-channel variants, the R channel is comparatively more informative.

- (c) **Attention Module Ablation:** Under the same LP-ViT architecture, data split, and training settings, we conduct comparative experiments on the attention-related components of LP-ViT. The tested configurations include: (i) the full model with both the dual-path LGQuery downsampling and positional encoding enabled; (ii) removing the local branch of LGQuery and retaining only the pooling branch; and (iii) removing positional encoding while keeping the complete LGQuery module. The results are reported in Table 10.

Table 10. Ablation results of the downsampling module.

Setting	Acc (%)	Macro-F1 (%)
LGQuery local	89.96	89.61
LGQuery pool	88.11	88.35
Positional Emb.	88.68	88.26

As shown in Table 10, the full configuration achieves the best recognition performance (Acc = 89.96%, macro-F1 = 89.61%). Removing the local branch of LGQuery or disabling positional encoding leads to a performance drop, indicating that the locally enhanced downsampling branch and explicit positional modeling both play important roles in improving the model's discriminative capability.

The above two ablation studies validate the contributions of the input representation and key architectural components to recognition performance. The following experiments further focus on the accuracy–efficiency trade-off of the student model for edge deployment, and optimize the distillation strategy and capacity configuration accordingly.

- (d) **Distillation temperature T melting:** Under the same teacher model, student LP-ViT architecture, data split, and training settings, we conduct comparative experiments by varying the distillation temperature T to evaluate how the temperature affects knowledge transfer effectiveness and edge-side inference efficiency. The results are reported in Table 11.

Table 11. Ablation results of the distillation temperature T.

Indicator	T = 1	T = 2	T = 3	T = 4	T = 5
Teacher inference time (s)	0.1970	0.1979	0.2240	0.1750	0.2049
Student inference time (s)	0.0160	0.0192	0.0122	0.0105	0.0152
Speedup	12.35×	10.33×	18.29×	16.70×	13.48×
Test Accuracy (%)	84.58	85.12	89.06	86.13	84.75

As shown in Table 11, T = 3 provides the best trade-off between recognition accuracy and inference efficiency for the student LP-ViT (accuracy = 89.06%, inference latency = 0.0122 s). When (T) is set too low or too high, the student model's accuracy decreases, indicating that the distillation temperature should balance the separability of soft targets and the expressiveness of inter-class relationships.

- (e) Ablation of the number of attention heads: Under the same teacher model, and with all other student LP-ViT configurations, data splits, and training settings held constant, we conduct comparative experiments by varying the number of attention heads to evaluate the effect of head count on the student model's recognition performance and edge-side inference cost. The results are reported in Table 12.

Table 12. Ablation results of the number of attention heads.

Indicator	num_heads = 4	num_heads = 2	num_heads = 1
Inference time (ms/sample)	5.5513	6.4540	6.1514
Parameter quantity (M)	0.37	0.37	0.37
Memory usage (MB)	56.88	53.74	52.73
Accuracy(%)	89.96	85.23	83.14

As shown in Table 12, num_heads = 4 provides the best trade-off between recognition accuracy and edge-side cost (accuracy = 0.8996, inference latency = 5.5513 ms/sample). When the number of heads is reduced to two or one, recognition performance drops noticeably, indicating that an overly small multi-head configuration constrains the student model's representational capacity.

- (f) (Network Depth Ablation: Under the same teacher model, and with all other student LP-ViT configurations, data splits, and training settings held constant, we conduct comparative experiments by varying the student network depth to assess the impact of depth on recognition performance and edge-side inference cost. The results are reported in Table 13.

Table 13. Ablation results of network depth.

Indicator	num_layers = 6	num_layers = 4	num_layers = 2	num_layers = 1
Inference time (ms/sample)	15.42	5.55	4.50	3.50
Parameter quantity (M)	0.570	0.37	0.17	0.12
Memory usage (MB)	72.15	56.88	51.25	48.72
Accuracy	0.8813	0.8796	0.8466	0.8183

As shown in Table 13, the 4-layer configuration provides the best overall balance between recognition performance and deployment cost (accuracy = 0.8996, inference latency = 5.55 ms/sample, and 0.37 M parameters). Although the 6-layer model achieves slightly higher accuracy, it incurs noticeably higher inference latency and memory usage. In contrast, the 2-layer and 1-layer models are faster, but their recognition performance degrades substantially.

4. Discussion

The main advances of this study can be summarized at three levels, namely task formulation, representation construction, and deployable method design. First, at the task level, we extend online Φ OTDR monitoring from simply determining whether anomalous disturbances are present to assessing the threat level. Specifically, we use four distance condition labels, 0 M, 5 M, 10 M, and 15 M, to describe the spatial proximity between events and the pipeline. In this way, the model outputs correspond directly to distance stratification, which is a key discriminative dimension in practice. Second, at the representation level, we fuse information from multiple sensing points into an RGB time–frequency image representation. This design jointly encodes time–frequency characteristics and spatial difference cues, enabling the model to learn spatial patterns associated with distance variations. Third, at the methodological level, we reduce the computational and memory overhead of a standard ViT by proposing LP ViT, and we further apply knowledge distillation to obtain a more compact student model while maintaining recognition performance. Concretely, the number of parameters is reduced from 85.65 M to 2.88 M, which corresponds to a 96.6 percent reduction. GPU memory usage decreases from 1368.71 MB to 95.6 MB, which corresponds to a 93 percent reduction. Training time is shortened from 42 s per epoch to 11.5 s per epoch. The distilled student model further reduces the parameter count to 0.37 M and achieves an inference latency of 5.5 ms per sample. Moreover, we conduct on-device validation on a Jetson TX2 and keep peak memory usage within 35 MB. This provides a reusable implementation and validation pathway for edge-oriented real-time threat level recognition.

From an engineering perspective, these threat level outputs can support response prioritization and guide verification strategies in online monitoring. When multiple alarms occur at the same time, higher-level events that are closer to critical pipeline segments can be reviewed first. Verification intensity and response deadlines can then be allocated according to the threat level, for example, by performing faster and more thorough checks for high-priority events while keeping lower-priority events under continued observation. This makes the disposal workflow more actionable. This stratification also enables clearer quantitative evaluation. Process-oriented metrics, such as the false alert notification rate, the number of events escalated to manual verification, repeated verification counts, and the average handling delay, can be summarized. Combined with labor and dispatching costs, these metrics can be used to assess how the grading strategy affects maintenance workload and operational costs.

Due to confidentiality constraints on field operation logs and cost data, we cannot disclose operational indicators such as dispatch frequency, response delay, or economic cost. To still provide a computable quantitative description of the potential load reduction achieved by the grading strategy, we introduce an offline evaluation framework based on an alarm queue. Let N denote the total number of alarms generated within a monitoring period. Under resource constraints, we assume that the operation team can verify at most the top K alarms, ranked by the score S or the predicted threat level. The theoretical reduction in manual verification workload is then $1 - K/N$. Based on this formulation, we define additional metrics to characterize the trade-off between reducing unnecessary checks and avoiding missed handling of high-risk events. These metrics include Top K high-risk coverage, defined as the recall of high-level events, such as 0 M and 5 M, within the top K alarms, the high-risk prioritization rate, and an upper bound on waiting time determined by the queue length and the processing rate. Together, these metrics provide an offline interface for evaluating the grading strategy and for guiding future optimization of thresholds and resource allocation in operational settings.

In addition, for pipeline sections near water sources, ecologically sensitive areas, or densely populated regions, the grading results help allocate response resources to events that are more likely to pose substantive risks. This can improve response timeliness and reduce potential environmental impacts. When combined with response feedback, the grading outputs also support a closed-loop workflow of continuous monitoring, event logging, and response feedback. This workflow enables long-term tracking of event time, location, threat level, and handling outcomes. Such structured records facilitate statistical analysis and post-incident auditing. They also provide a more systematic basis for regulatory inspections and for assessing infrastructure resilience.

This study has several limitations. The experimental data were collected from a limited set of scenarios within a single test environment, so both the sample scale and the coverage of environmental conditions are constrained. In particular, the dataset does not systematically capture signal variations caused by different soil types, pipeline materials, burial depths, backfill conditions, or more complex background noise. Examples of such noise include traffic, wind, rain, and pump station vibrations. Therefore, the reported results primarily reflect the effectiveness and stability of the proposed method under the current acquisition conditions. Robustness under more complex or highly variable field environments still requires validation using larger and more diverse datasets.

In addition, the correspondence between the proposed threat levels and engineering risk requires further clarification. In this work, threat levels are defined operationally using relative distance, which supports rapid stratification. However, with richer field data and stronger risk assessment priors, it will be necessary to establish a more interpretable mapping between the grading scheme and risk-relevant quantities. Examples include potential damage likelihood, expected response cost, and safety criticality.

Building on the above, future work will proceed along three directions. First, we will expand cross-site data collection and evaluation across operating conditions. This will cover diverse soil types and pipeline materials, different construction and excitation sources, and more complex background noise environments. The goal is to systematically assess model generalization under realistic deployment conditions. We will also conduct noise-oriented sensitivity analyses by constructing evaluation settings with multiple noise types and intensity levels. For each setting, we will quantify changes in accuracy, macro-F1, and confusion matrices. We will further report performance degradation trends as noise intensity increases to characterize the stability boundary of the model in complex noise backgrounds. Second, we will develop an operations-oriented metric system that better matches practical workflows. Examples include reductions in false trigger notifications under the grading strategy, fewer repeated verifications, and shorter average handling delays. These metrics will be used to estimate potential savings in maintenance costs and improvements in response efficiency. Third, we will investigate rapid calibration and transfer strategies for new scenarios. We will also link grading outputs with event logging and tracking mechanisms to improve system sustainability and maintainability during long-term online operation.

5. Conclusions

This study addresses challenges in Φ -OTDR pipeline monitoring systems, including weak signal recognition, high environmental noise, and real-time requirements, by proposing an efficient early warning scheme tailored for engineering applications. A correlation framework maps excavator distances to pipeline threat levels, converting complex vibration signals into quantifiable risk indicators. Multi-dimensional feature fusion and modality conversion mechanisms effectively suppress noise and reconstruct transient vibration features. To reduce the computational burden of traditional transformer models, a local-global

hybrid attention architecture is developed, reducing parameters by 96.6% from 85.65 M to 2.88 M. Using a teacher–student framework for model compression, parameters are further optimized to 0.37 M. The model achieves a 5.5 ms inference latency on a Jetson TX2 edge device, ensuring rapid processing from signal acquisition to threat detection, while balancing false alarm rates and response speed for real-time applications.

Author Contributions: Conceptualization, Y.Z. and H.Z.; methodology, Y.Z.; software, Y.Z. and J.Y.; validation, Y.Z. and C.S.; formal analysis, Y.Z.; investigation, Y.Z.; resources, Y.Z.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.; visualization, Y.Z.; supervision, J.Z. and J.W.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Nos. 52376159, 52474064), the Opening Project of the State Key Laboratory of Shale Oil and Gas Enrichment Mechanisms and Effective Development, and Frontier Interdisciplinary Exploration Research Program of China University of Petroleum, Beijing (Grant No. 2462024XKQY005).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to commercial confidentiality.

Conflicts of Interest: The authors declare no conflicts of interest. The funders 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.

Abbreviations

The following abbreviations are used in this manuscript:

Φ -OTDR	Phase-Sensitive Optical Time Domain Reflectometry
AOM	Acousto-Optic Modulator
EDFA	Erbium-Doped Fiber Amplifier
DAQ	Data Acquisition Card
RPP	Relative Position Patch
SPE	Sinusoidal Positional Encoding
DPCM	Dual-Path Compression Module
L-MHA	Lightweight Multi-Head Attention
STFT	Short-Time Fourier Transform
ISTFT	Inverse Short-Time Fourier Transform

References

1. Ramírez-Camacho, J.G.; Carbone, F.; Pastor, E.; Bubbico, R.; Casal, J. Assessing the consequences of pipeline accidents to support land-use planning. *Saf. Sci.* **2017**, *97*, 34–42. [\[CrossRef\]](#)
2. Huang, Y.; Qin, G.; Yang, M. A risk-based approach to inspection planning for pipelines considering the coupling effect of corrosion and dents. *Process Saf. Environ. Prot.* **2023**, *180*, 588–600. [\[CrossRef\]](#)
3. Khan, F.; Thodi, P.; Imtiaz, S.; Abbassi, R. Real-time monitoring and management of offshore process system integrity. *Curr. Opin. Chem. Eng.* **2016**, *14*, 61–71. [\[CrossRef\]](#)
4. Rao, Y.; Wang, Z.; Wu, H.; Ran, Z.; Han, B. Recent Advances in Phase-Sensitive Optical Time Domain Reflectometry (Φ -OTDR). *Photonic Sens.* **2021**, *11*, 1–30. [\[CrossRef\]](#)
5. Wang, Z.N.; Zeng, J.J.; Li, J.; Fan, M.Q.; Wu, H.; Peng, F.; Rao, Y.J. Phase-sensitive optical time-domain reflectometry with counterpumping fiber Brillouin amplification. *Opt. Lett.* **2014**, *39*, 4313–4316. [\[CrossRef\]](#)
6. Peng, F.; Duan, N.; Rao, Y.J.; Li, J. Ultra-long high-sensitivity Φ -OTDR for high spatial resolution intrusion detection of pipelines. *Opt. Express* **2014**, *22*, 13804–13810. [\[CrossRef\]](#)
7. Liu, X.; Wang, Y.; Jin, B.; Ying, Q.; Wang, D.; Wang, Y. Realtime distributed oil/gas pipeline security pre warning system based on Φ -OTDR. In Proceedings of the Progress in Electromagnetics Research Symposium (PIERS-FALL), Singapore, 19–22 November 2017; pp. 1717–1719.

8. Wang, Z.; Zhang, J.; Zhang, L.; Li, S.; He, Y. Novel railway-subgrade vibration monitoring technology using phase-sensitive OTDR. In Proceedings of the 25th Optical Fiber Sensors Conference (OFS), Jeju, Republic of Korea, 24–28 April 2017; pp. 1–4.
9. Jiang, Y.; Wang, Y.; Cao, P.; Safari, M.; Thompson, J.; Haas, H. Robust and low-complexity timing synchronization for DCO-OFDM LiFi systems. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 53–65. [[CrossRef](#)]
10. Zhou, J.; Pan, Z.; Ye, Q.; Cai, H.; Qu, R.; Fang, Z. Characteristics and explanations of interference fading of a Φ -OTDR with a multi-frequency source. *J. Light. Technol.* **2013**, *31*, 2947–2954. [[CrossRef](#)]
11. Rayo, M.F.; Moffatt-Bruce, S.D. Alarm system management: Evidence-based guidance encouraging direct measurement of informativeness to improve alarm response. *BMJ Qual. Saf.* **2015**, *24*, 282–286. [[CrossRef](#)]
12. Chang, Y.; Khan, F.; Ahmed, S. A risk-based approach to design warning system for processing facilities. *Process Saf. Environ. Prot.* **2011**, *89*, 310–316. [[CrossRef](#)]
13. Ma, P.; Liu, K.; Jiang, J.; Li, Z.; Li, P.; Liu, T. Probabilistic event discrimination algorithm for fiber optic perimeter security systems. *J. Light. Technol.* **2018**, *36*, 2069–2075. [[CrossRef](#)]
14. Tejedor, J.; Martins, H.F.; Piote, D.; Macias-Guarasa, J.; Pastor-Graells, J.; Martin-Lopez, S.; Corredera, P.; De Smet, F.; Postvoll, W.; Gonzalez-Herraez, M. Toward prevention of pipeline integrity threats using a smart fiber-optic surveillance system. *J. Light. Technol.* **2016**, *34*, 4445–4453. [[CrossRef](#)]
15. Lyu, C.; Niu, Z.; Tian, J.; Jin, J.; Yang, J.; Ge, C. Identification of intrusion events based on distributed optical fiber sensing in complex environment. *IEEE Internet Things J.* **2022**, *9*, 24212–24220. [[CrossRef](#)]
16. Wang, B.; Pi, S.; Sun, Q.; Jia, B. Improved wavelet packet classification algorithm for vibrational intrusions in distributed fiber-optic monitoring systems. *Opt. Eng.* **2015**, *54*, 055104. [[CrossRef](#)]
17. Huang, X.; Zhang, H.; Liu, K.; Liu, T.; Wang, Y.; Ma, C. Hybrid feature extraction-based intrusion discrimination in optical fiber perimeter security system. *IEEE Photonics J.* **2016**, *9*, 1–12. [[CrossRef](#)]
18. Wu, H.; Li, Z.; Li, P.; Liu, K.; Liu, T. One-dimensional CNN-based intelligent recognition of vibrations in pipeline monitoring with DAS. *J. Light. Technol.* **2019**, *37*, 4359–4366. [[CrossRef](#)]
19. Liu, L.; Sun, W.; Zhou, Y.; Li, Y.; Zheng, J.; Ren, B. Security event classification method for fiber-optic perimeter security system based on optimized incremental support vector machine. In *Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 484, pp. 595–603.
20. Liu, K.; Liu, T.; Li, P.; Li, Z.; Jiang, J.; Feng, X. A high-efficiency multiple events discrimination method in optical fiber perimeter security system. *J. Light. Technol.* **2015**, *33*, 4885–4890. [[CrossRef](#)]
21. Kye, M.; Lee, S.-M.; Choi, K.-M.; Lee, Y.; Park, K.-Y. A surveillance system of fiber-optic cables with multi-channel DAS and CNN. *IEEE Photonics Technol. Lett.* **2021**, *33*, 753–756. [[CrossRef](#)]
22. Chen, X.; Xu, C. Disturbance pattern recognition based on an ALSTM in a long-distance Φ -OTDR sensing system. *Microw. Opt. Technol. Lett.* **2020**, *62*, 168–175. [[CrossRef](#)]
23. Wu, H.; Yang, S.; Liu, X.; Xu, C.; Lu, H.; Wang, C.; Qin, K.; Wang, Z.; Rao, Y.; Olaribigbe, A.O. Simultaneous extraction of multi-scale structural features and the sequential information with an end-to-end mCNN-HMM combined model for DAS. *J. Light. Technol.* **2021**, *39*, 6606–6616. [[CrossRef](#)]
24. Wu, H.; Yang, M.; Yang, S.; Lu, H.; Wang, C.; Rao, Y. A novel DAS signal recognition method based on spatiotemporal information extraction with 1DCNNs-BiLSTM network. *IEEE Access* **2020**, *8*, 119448–119457. [[CrossRef](#)]
25. National People's Congress of the People's Republic of China. *Law of the People's Republic of China on the Protection of Petroleum and Natural Gas Pipelines*; National People's Congress: Beijing, China, 2010.
26. ASME B31.8S; Managing System Integrity of Gas Pipelines. American Society of Mechanical Engineers: New York, NY, USA, 2022.

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.