Pipeline Safety Early Warning by Multifeature-Fusion CNN and LightGBM Analysis of Signals From Distributed Optical Fiber Sensors

Yiyuan Yang, Haifeng Zhang, and Yi Li

Abstract—Energy pipelines are the backbones of global energy systems. Monitoring their safety and automatically identifying and locating third-party damage events are crucial to energy supply. However, most traditional methods lack in-depth consideration of distributed fiber signals and have not been tested on real-world long-distance pipelines, making it difficult to deploy them in operating long-distance pipelines. In this study, we utilize a novel real-time machine-learning method based on phase-sensitive optical time domain reflectometer technology to monitor the safety of oil and gas pipelines. Specifically, we build a multifeature-fusion convolutional neural network and LightGBM fusion model based on two novel complementary spatiotemporal features. The method was applied to a large amount of data collected from real-world oil–gas transportation pipelines of the China National Petroleum Corporation. The proposed method could accurately locate and identify third-party damage events in real-time under conditions of strong noise and various types of system hardware, and could effectively handle signal drift in the time and space dimensions. Our methodology has been deployed at real long-distance energy pipeline sites and our work will contribute to energy pipeline safety and energy supply security. Furthermore, the proposed solution could be generalized to other fields, such as industrial inspection, measurement, and monitoring.

Index Terms—Distributed optical fiber sensor, industrial signal processing and monitoring, lightGBM, multifeature fusion convolutional neural network (MFCNN), pipeline safety early warning (PSEW).

I. INTRODUCTION

OIL and gas pipelines are widely used in the field of energy transportation because of their low cost, small size, and rapid construction. According to the latest statistics from the Central Intelligence Agency (CIA), the total length of long-distance oil and gas transportation pipelines

worldwide is 3.55×10^6 km and is increasing by approximately 30000 km per year. However, at the same time, the accident rates associated with energy pipelines remain high. For example, the rate is 0.25 accidents per year per 1000 km in Europe, 0.5 accidents per year per 1000 km in the United States, and three accidents per year per 1000 km in China. These accidents may cause oil and gas leakages or even explosions, considerable indirect economic losses, environmental pollution, energy crises, personnel safety issues, and extremely negative publicity [1].

Most pipelines are buried underground to reduce their floor space, which makes it difficult to observe their safety aspects directly [2]. Moreover, most pipelines are laid through complex environments such as farmland, deserts, hills, and other remote areas, making them highly unsuitable for manual inspection. Current approaches of pipeline safety inspection include manual patrolling, unmanned aerial vehicle (UAV) patrolling, and camera-assisted method at key locations, but all of them cannot be used for automated, distributed, and real-time pipeline inspections [3]. With the increasing lengths of pipelines, these traditional and high-cost monitoring approaches have become a major obstacle to the development of modern and intelligent pipeline transportation systems. Consequently, using intelligent early warning algorithms to replace traditional manual monitoring could be of great benefit to the energy pipeline industry and its systems.

Recently, sophisticated sensors, advanced artificial intelligence (AI) algorithms, cloud computing technology, and big data systems have made it possible to apply intelligent pipeline safety early warning (PSEW) systems in the industry. Optical fiber sensors are currently considered useful by researchers because of their weak radiation, long-distance laying ability, high precision, noncontact measurement characteristics, and good real-time performance [4], [5]. Optical fibers are often already used in pipeline cables for the company's internal communication and data transmission purposes [6]. Researchers have attempted to utilize these in a distributed PSEW system through phase-sensitive optical time domain reflectometer (ϕ-OTDR) technology without any increase in operating cost [7], [8], which can meet the typical industrial requirements of spatial resolution of 5–20 m. Compared to other optical fiber systems, such as Brillouin optical time
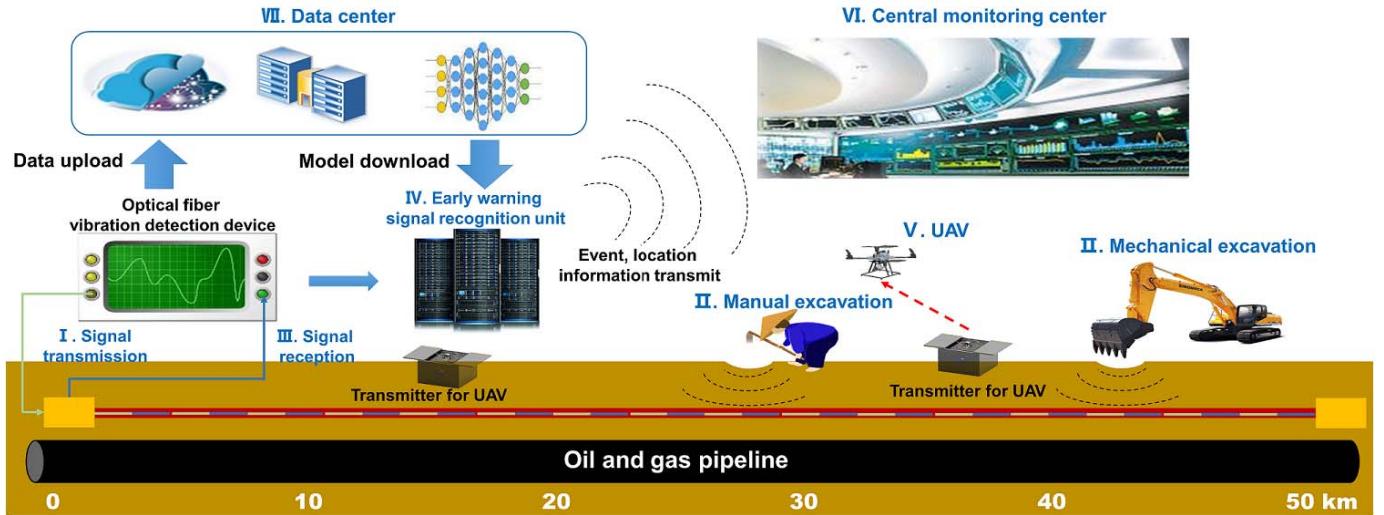
Fig. 1. Structure of an optical-fiber-based PSEW system [9]. The signal transmitter sends a narrow-pulse signal into the optical fiber through a coupler (I). When a dangerous event occurs (II), the signal carries its information back to the signal receiver (III). Then, the early warning unit immediately processes and recognizes the signal (IV). If the prediction is of a damage event, the system notifies the UAV closest to the event to automatically gather evidence and issue a warning immediately (V). Detailed information is sent to the central monitoring center for filing (VI). Also, the system periodically uploads information to the data center to optimize and update the model and automatically download the latest model to the early warning unit (VII).

domain analysis (BOTDA) and fiber Bragg grating, ϕ-OTDR has the advantages of sensitive, relatively high signal energy, and support long-distance measurements. Besides, compared to BOTDA, it supports single-ended access, i.e., when a break in the fiber occurs, the part of the system before the break can still work and the location of the break can be quickly located. Therefore, ϕ-OTDR is currently the most common technology used for PSEW system [3]. There has also been concurrent development of big data and data storage technology that can provide high-quality data resources for data-driven applications. Through distributed computing, cloud computing technology can be used to break down huge data-processing procedures into multiple small programs. Their results can then be processed and analyzed through a system of multiple servers, which can process tens of thousands of data points in a very short time. At the same time, AI algorithms can build models with powerful abilities to approximate real-time PSEW and spatiotemporal positioning of third-party damage events. The novel optical-fiber-based oil–gas PSEW system shown in Fig. 1 is such an intelligent system that recognizes and locates dangerous events, issues early warnings, conducts on-site inspections, and records data in real-time.

Although optical fiber sensors have the advantages of good real-time performance and easy installation [10], their internal signals are susceptible to fluctuations because of environmental influences [11]. Specifically, they have strong noise, weak signals, signal jitter, and the problem of signal drift over time and space [12], which places higher demands on algorithms that recognize optical fiber sensor signals. Algorithms that have been researched for recognizing and locating third-party damage events in PSEW systems are mainly of two types: 1) *Traditional signal-processing methods*, such as wavelet decomposition (WD), wavelet packet decomposition (WPD), mel-frequency cepstral coefficients (MFCC),

domain power spectrum, and other methods in the frequency domain [13]–[16]. For example, Tanimola and Hill [13] proposed using distributed temperature sensing and distributed acoustic sensing, and Ting, *et al*. [14] used WD and WPD to extract features. Tabi Fouda *et al*. [15] presented an estimation method based on a frequency-domain power spectrum, and Zhang *et al*. [16] applied MFCCs to extract features; 2) there are *machine-learning and deep-learning methods*, such as Bayesian networks and their variants, the stochastic configuration network (SCN) and its variants, hidden Markov models (HMMs), convolutional neural networks (CNNs), and probabilistic neural networks (PNNs) [17]–[30]. For instance, Kabir *et al*. [17] and Guo *et al*. [18] used a Bayesian network and variants for PSEW. Sheng *et al*. [19] updated the SCN proposed by Wang and Li [20] based on truncation singular value decomposition, and called it TSVD-SCN. Wu *et al*. [21] applied an HMM to extract event areas and judge event categories. There are also other kinds of deep-learning methods that can be used for PSEW [22]–[26]. In particular, Shi *et al*. [23], Wu *et al*. [27], Yang *et al*. [28], and Wu *et al*. [29] proposed modified CNNs and variants, whereas Kong *et al*. [30] applied a PNN to solve this problem. Besides, Yang *et al*. [36] used bilinear CNN and LightGBM, and this article is an extension of it.

Nevertheless, PSEW algorithms for use with distributed optical fiber remain largely unexplored. Although the methods described above have demonstrated good performance using optical fiber sensor data in ideal experiments, it remains difficult to successfully apply them to actual industrial processes. The main reason is the difference between the data distributions of samples obtained under experimental and real conditions. Reducing and eliminating the negative impacts caused by such differences are key to generalizing predictive models derived from experimental data to real industrial applications [9].

Specifically, the following problems with PSEW systems remain: 1) the spatiotemporal features of sensor signals consistently change, making them difficult to process with a single algorithm or a single feature [2]; 2) strong noise, weak signals, and signal jitter, and the problem of signal drift over time, make it difficult for an algorithm trained under ideal conditions to fit complex environments in real industrial applications [10], [36]; 3) a low-frequency signal can be processed at higher speed with a cheaper solution, but because of its transmission of less information per unit time, a higher requirement is imposed on the algorithm [9]; 4) most of the existing methods described above do not consider the characteristics of distributed sensors, only extract features in the time domain, and do not fuse the information of the spatiotemporal dimensions [3], [5], [12]; 5) existing research has not been validated for use with long-distance pipelines already in service, which makes them unconvincing for use with real site data characterized by strong noise, weak signals, and drift problems [3], [13].

This article is motivated by strong industrial demand. We expect to employ advanced ML techniques to improve the performance and robustness of PSEW algorithms in identifying and locating damage events in real time in various environments, and to reduce the cost of deployment. Our contributions are summarized as follows.

1) We describe a novel application scenario (i.e., the identification and location of third-party damage events in a long-distance PSEW system) for different ML techniques, including MFCNNs and LightGBM.
2) We propose two dimensionless and complementary features of the time domain that are based on spatiotemporal information in distributed optical fiber sensor signals.
3) We present a novel multifeature-fusion ML model to achieve a robust PSEW system and effectively mitigate the negative impacts of the difference between training and testing samples on localization and recognition accuracy.
4) We have deployed our equipment and conducted experiments in two operating long-distance oil–gas pipelines to collect data from different hardware under various environmental conditions and during multiple dangerous events.
5) We evaluate the effectiveness of our method with other baselines using different experimental datasets and find that our model is more adaptable to complex environments and more scalable to hardware than other baselines. It also demonstrates good real-time performance, high recognition accuracy, and positioning precision.

The remainder of this article is organized as follows. The concepts of the ML techniques used are presented in Section II, including CNNs, MFCNNs, and LightGBM. Section III introduces the details of a pipeline safety early warning experimental facility and our data acquisition method. A complete approach for locating and identifying third-party pipeline damage events in real-time is described in Section IV. Comprehensive case studies, system deployment, and a discussion are presented in Section V. A summary of this article and future research direction are given in Section VI.

II. BACKGROUNDS TO RELATED MACHINE-LEARNING TECHNIQUES

A. Convolutional Neural Networks

For the identification and localization of events that compromise pipeline safety based on distributed optical fiber signals with high-dimensional input, it is not practical to apply fully connected neural networks to generate a classifier, since a high number of weight parameters is necessary for each neuron. Fortunately, CNNs provide a solution to reduce the parameter number. The convolutional neuron shares the same weights as each spatial location (i, j). Therefore, the parameter number is greatly reduced. The output $o_{i,j}$ of a convolutional layer with location (i, j) is as follows:

$$o_{i,j} = \sigma \left((W_c \otimes X)_{i,j} + b_c \right) \tag{1}$$

where W_c is the learned weight of the convolutional kernel, which is a matrix, b_c is the bias, $\otimes$ is the convolutional operation, and $\sigma(\cdot)$ is the activation function. Common activation functions include the rectified linear unit (Relu) and Sigmoid. Besides, we utilize batch normalization (BN) layers to normalize the data before the activation function to resolve the impacts of shifting and increasing input data [31]. After the convolutional layer, it is common to add a pooling layer to reduce the data dimension. The most common pooling layer includes max pooling and average pooling.

A fully connected layer is required to generate the classification stage after the convolutional, BN, and pooling layers. The mathematical representation of the fully connected layer is

$$o_{i,j} = \sigma \left((W_f * X)_{i,j} + b_f \right) \tag{2}$$

where W_f is the learned weight matrix, and b_f is the bias. The activation function for multiclassification is Softmax

$$P(C_i|X) = \frac{e^{V_i(X)}}{\sum_{i=1}^{C} e^{V_i(X)}} \tag{3}$$

where $V_i(X)$ is the ith input data of the Softmax layer with instance X, and C is the number of categories. $P(C_i|X)$ is the probability of class i with instance X. If $P(C_i|X) \geq P(C_j|X)$, then the unknown instance will be identified as class i and vice versa.

In addition, the weight matrix above W_c, W_f and the bias b_c, b_f in the CNN and fully connected layer are updated through minimization of the loss between the ground truth $\{y_i^1, y_i^2, \ldots, y_i^C\}$ and the prediction $\{\hat{y}_i^1, \hat{y}_i^2, \ldots, \hat{y}_i^C\}$, for N training instances $i = 1, \ldots, N$. The categorical cross-entropy loss is used as the loss function for multiclassification, shown as follows:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i^1 log(\hat{y}_i^1) + y_i^2 log(\hat{y}_i^2) + \ldots + y_i^C log(\hat{y}_i^C) \right]. \tag{4}$$

The categorical cross-entropy loss has fast convergence rates and is numerically stable when coupled with Softmax.

Besides, we need to use a suitable optimization algorithm to minimize the loss function. Common optimizers include root mean square propagation (RMSprop), adaptive moment estimation (Adam), stochastic gradient descent (SGD), adaptive gradient algorithm (AdaGrad), an extension of AdaGrad (AdaDelta), and Nesterov-accelerated adaptive moment estimation (Nadam) [32]. Specifically, Adam has the advantages of fast convergence and easy reference adjustment [33].

B. Multifeatures Fusion CNN

Although CNNs have good function approximation capability, proper signal transformation and feature fusion are equally important. Therefore, the multifeature fusion technique is introduced to acquire complementary advantages between different features. Taking two features as an example, the results of two features in the same labeled dataset should be similar. So, in addition to the loss between the label and the predicted values of each of the two features, the penalty between the predicted label of the two features needs to be considered. Without loss of generality, let p and q denote two probability distribution vectors of the same length D, i.e., the number of categories, and we define the penalty function $L(p, q)$ by employing the mean squared error, as shown as follows:

$$L(p, q) = \frac{1}{D} \| p - q \|_2^2. \tag{5}$$

So, we consider a training set $\{x_1, x_2, \ldots, x_N\}$ where the number of training samples is N. Let x_i^1 and x_i^2 denote the two extracted features of x_i, for N training instances $i = 1, \ldots, N$, respectively. Let w^1 and w^2 denote the learned parameters of two branches of the CNN. Thus, the loss function of the multifeature fusion model can be written as follows:

$$L_{\text{all}} = \frac{\lambda_1}{N} \sum_{i=1}^{N} L\left(f_{w^1}(x_i^1), y_i\right) + \frac{\lambda_2}{N} \sum_{i=1}^{N} L\left(f_{w^2}(x_i^2), y_i\right)$$
$$+ \frac{\lambda_3}{N} \sum_{i=1}^{N} L\left(f_{w^1}(x_i^1), f_{w^2}(x_i^2)\right) + \lambda_4 \| w^1 \|_2^2 + \lambda_5 \| w^2 \|_2^2 \tag{6}$$

where y_i denotes the ground truth of the ith training sample, and the set $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ denotes the hyperparameters that control the weights of each part of the model and $L2$ regularization, respectively.

Then, we use an optimization algorithm, such as Adam, to minimize the above loss function, and we can get the optimal w_{opt}^1 and w_{opt}^2 values to use for prediction.

C. LightGBM

LightGBM is a novel gradient boosting decision tree (GBDT) algorithm proposed by Microsoft in 2017 [34]. Its advantages include automatic feature processing, avoidance of overfitting, efficient training, distributed support, fast running speed, and low memory overhead, making it more appropriate and faster in industrial practice than other algorithms. Specifically, LightGBM uses gradient-based one-side sampling (GOSS) to exclude most of the samples with small gradients and calculates information gain with other data. Besides, unlike another common GBDT called

XGBoost, LightGBM applies a leaf-wise generation strategy instead of the traditional level (depth)-wise approach to reduce the number of training data. Furthermore, LightGBM employs exclusive feature bundling (EFB) to turn many high-dimensional mutually exclusive features into low-dimensional dense features, which avoids computation of redundant features, especially unnecessary zero-valued ones. Also, LightGBM uses histogram computing and parallel learning methods to optimize the calculation speed.

III. EXPERIMENTAL FACILITY AND DATA ACQUISITION

A. Experimental Facility

As shown in Fig. 2(b), we use a redundant single-mode fiber from the cable installed at the same time as the pipeline for communication, and generate a light source using the Rayleigh scattering technique. The actual devices are shown in Fig. 2(a). In detail, the laser we used is a narrow linewidth laser from NKT Co. with an output power of 40 mW. The optical fiber is a single mode (Class B) dispersion-shifted fiber with a central wavelength of 1550 nm and an average power loss of < 0.26 dB/km. The data acquisition card (DAC) is from NI, USA, and the server is from ThinkPad. The rest of the hardwares are custom-made by ourselves, and no other commercial systems were used.

A simplified schematic of the deployed system is shown in Fig. 2(c). It uses an ultranarrow linewidth light (UNLL) source with a center wavelength of 1550 nm. An acoustic optic modulator (AOM) shift chops the continuous light into probe pulses and an erbium-doped fiber amplifier (EDFA) is deployed to compensate for light energy losses. The amplified probe pulses are injected into the optical fiber through a fiber circulator (FC). When the vibrations generated by third-party events propagate through the soil into the fiber, the length and refractive index of the fiber at the corresponding location will change because of the photoelastic effect, causing a change in the phase of the backward Rayleigh scattered light as well as the change of coherent interference result, thus capturing the disturbance information of the fiber. Meanwhile, the event can be located by the round-trip time of the light pulse between the detection site and the light source [3]. Next, Rayleigh backscattered light containing event information is directly routed to a photoelectric detector (PD). The intensity evolution over time is recorded by a DAC at frequencies of 5 MHz for the time dimension and 500 Hz for the spatial dimension and processed by a personal computer (PC) [35]. To adapt to the monitoring requirements of long-distance pipelines (> 50 km), we deploy a signal relay amplifier (SRA) every 25 km along the pipe for signal boosting, which solves the problem of the low signal-to-noise ratio (SNR) at the end of the optical fiber.

B. Data Acquisition

Our equipment was deployed in two oil–gas pipelines of the China National Petroleum Corporation (CNPC). The collected datasets are described in Table I. Case 1 in the Table I is a 48-km gas pipeline of the West-East Natural Gas Transmission Project Suzhou section, which will be discussed in Section V-B, and Case 2 is an 85-km oil pipeline of
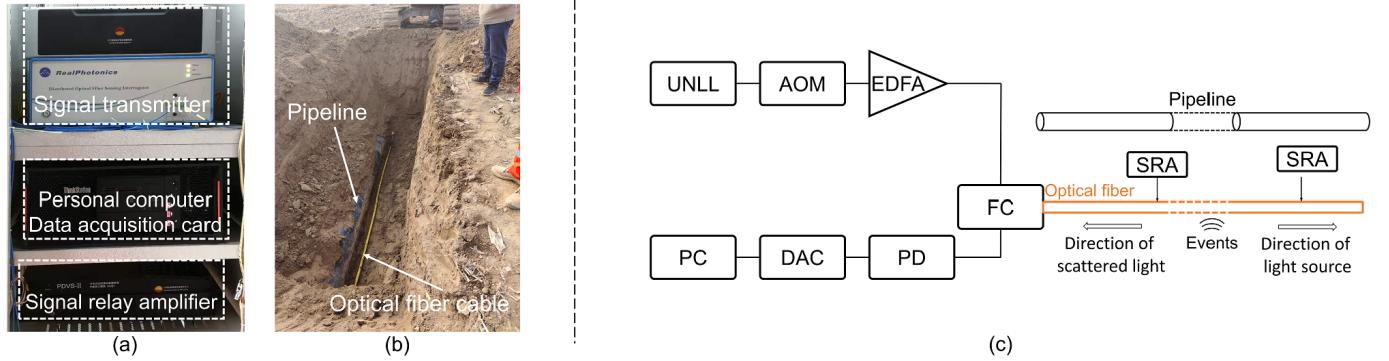
Fig. 2. PSEW system experimental facility. (a) Actual devices. (b) Relative positions of the optical fiber cable and pipeline, which are typically 20–30 cm apart. (c) Structural diagram of a deployed PSEW system.

TABLE I
DESCRIPTIONS OF THE COLLECTED DATASETS

Event	Label	Case 1		Case 2	
		100 Hz (#)	500 Hz (#)	100 Hz (#)	500 Hz (#)
Background noise	1	10,540	10,540	13,500	13,500
Manual excavation	2	16,980	12,060	3,920	3,600
Mechanical excavation	3	9,200	8,100	16,600	14,400
Vehicle driving	4	13,920	10,300	10,800	9,600

the National Pipeline Network Northern Pipeline Company Cangzhou section, which will be discussed in Section V-C. Validation experiments were conducted in each of the two pipelines. In contrast to ideal conditions, the collected signal was weak with strong noise and signal drift, which is unique to real-world long-distance pipelines. Moreover, to verify the universality and cost of the hardware, we tested 100- and 500-Hz signals, respectively. The lower the frequency, the lower the equipment price but the lower the data collection rate. Therefore, higher requirements were put forward to the algorithm. It made engineering sense to test the 500- and 100-Hz signals separately.

The categories of damaging events included *background noise* (no damaging events), *manual excavation* (potential oil theft by drilling of the pipe), *mechanical excavation* (third-party construction damaging the pipe), and *vehicle driving* (potential threat of heavy vehicles rolling over the pipe). Specifically, to investigate mechanical and manual excavations, we conducted a wide variety of experiments, as shown in Fig. 3. We performed mechanical excavations at positions 20, 10, and 5 m to the side of the pipeline, and directly above it [Figs. 3(a)–(d)], to verify the ability of our algorithm to recognize the distance of an event from the pipe. For manual excavations [Figs. 3(e)–(h)], we experimented with various scenarios, such as underground digging by spade and surface digging with by hoe, spade, and spade plus hoe to verify the robustness of our method in these cases. Then, we labeled the precise categories and spatiotemporal coordinates of each event.

There were two main difficulties in collecting the data: 1) *vast manpower consumption*. Professionally skilled technicians were required to travel along the pipeline to simulate intrusions in poor and complex environments. Additionally,

Fig. 3. Real scenarios used for data acquisition and experiments.

to acquire sufficiently diverse data, intrusions would need to be simulated at various locations in dozens of pipelines, which would cost over $100, 000, and 2) *high time consumption*. To verify the signal drift and robustness of the algorithm, we conducted long experiments of more than five months in real pipeline sites with harsh environments.

IV. METHODOLOGY

The proposed method first performs signal compensation to suppress the energy loss of the signal because of transmission in the fiber, then uses a novel feature generator to obtain the spatiotemporal characteristics of the two complementary signals. Finally, it applies the described MFCNN and LightGBM fusion model for localization and identification.
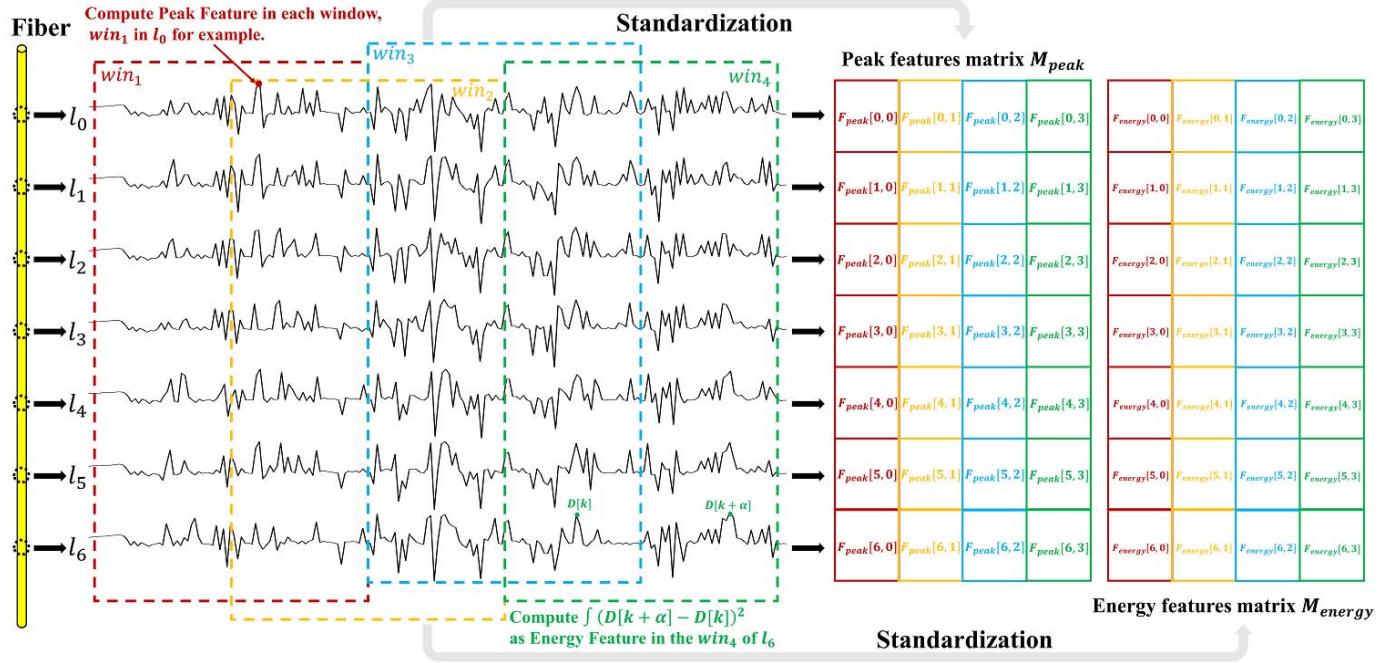
Fig. 4. Visualization example of the proposed feature extractor. The parameters in Algorithm 1 used in this instance are as follows. The length of sliding window N_{swin} is 60 and the sliding step N_{sstp} is 30, the total number of data in time domain T is 200, the number of considered observation points N_{cpt} is 7, and the number of the considered windows N_{cwin} is 4.

A. Signal Compensation

When Rayleigh backscattered light is transmitted through fiber, its energy decreases exponentially with propagation distance. More specifically, the peak power $P(l)$ with attenuation coefficient α is $P(l) = P_0 \, e^{-\alpha l}$, where l is the propagation distance and P_0 is the initial peak power. So, the Rayleigh scattering power is $P_R(l) = (P_0 e^{-\alpha l} S \alpha_s W v / 2)$, where α_s is the Rayleigh scattering coefficient, which usually ranges from 0.12 to 0.15 db/km, $S = (1/4)((\lambda/\pi nr))^2$ is the backscattered light power capture coefficient (with wavelength λ, fiber core refractive index n, and fiber mode field radius r), W is the pulsewidth of the light, and v is the speed of light in the fiber. When the signal is received, its power becomes $P_R(l) = (P_0 e^{-2\alpha l} S \alpha_s W v / 2)$. Therefore, we use SRAs at approximately 25-km intervals to handle signal attenuation. To ensure the universality of the algorithm for signals at different distances, we compensate for the segmented attenuation of the signals between boosters based on least-squared error with an exponential function.

B. Feature Generator

To allow supervisors and technicians at stations along the pipeline to visually monitor the situation for more intuitive analysis, we need to extract the feature manually to ensure its interpretability instead of using the visualization results directly from the middle layer of the model [36]. Also, we need to guarantee the real-time performance of the feature generator, which requires us to discard traditional frequency domain methods, such as the time-consuming Fourier transform and even fast Fourier transform. Moreover, to improve the robustness of the features, we need to build dimensionless features.

Based on the above, we propose two novel features that are both calculated in the time domain and can represent low- and high-frequency spatiotemporal information. Specifically, the peak feature represents high-frequency and instantaneous information, and the energy feature describes low-frequency and continuous information. Algorithm 1 shows the calculation method in detail. Besides, Fig. 4 visualizes an example of the calculation process of the feature extractor.

C. Event Recognizer With MFCNN and LightGBM Fusion Model

In the event recognizer, we fuse MFCNN and LightGBM to perform spatiotemporal localization and recognition. Specifically, we use MFCNN for initial feature extraction to build the feature vectors, and then apply LightGBM for more accurate localization and recognition. The specification of the model we used is summarized in Table II.

1) The standardized peak features and energy features generated by the feature generator are used as inputs to the two branches of the MFCNN. We take the number of considered observation points N_{cpt} as 7, and the number of considered windows N_{cwin} as 100, i.e., the input size of each branch is 7×100, which corresponds to 20 s of 500-Hz signals or 100 s of 100-Hz signals, and 120 m of pipeline.

2) The network is trained by allocating a random 60% of the total dataset as a training set. Specifically, we use sparse category cross entropy with four events as our base loss function, and the total loss is given by (6). Also, using Adam as the optimizer, we monitor the loss values of the validation set for hyperparameter tuning.

TABLE II
MODEL SPECIFICATION

Peak model			Energy model		
Layer	Shape	Activation function	Layer	Shape	Activation function
BN	(1,7,100)	-	BN	(1,7,100)	-
Conv1	(16,7,100)	Relu	Conv1	(16,7,100)	Relu
Max-pool1	(16,4,50)	Maxpool	Max-pool1	(16,4,50)	Maxpool
Conv2	(32,4,50)	Relu	Conv2	(32,4,50)	Relu
Max-pool2	(32,2,25)	Maxpool	Max-pool2	(32,2,25)	Maxpool
Conv3	(64,2,25)	Relu	Conv3	(64,2,25)	Relu
Max-pool3	(64,1,7)	Maxpool	Max-pool3	(64,1,7)	Maxpool
Flatten	(448)	-	Flatten	(448)	-
Fc1+BN	(64)	Relu	Fc1+BN	(64)	Relu
Dropout [37]	(64)	-	Dropout	(64)	-
Fc2	(4)	Softmax	Fc2	(4)	Softmax
LightGBM					
max_depth	num_leaves	min_data_in_leaf	boosting	learning_rate	lambda_l1
4	10	51	gbdt	0.003	0.1

Algorithm 1 Processing of Feature Generator

1: **Input:** Raw data X, Historical data X_{old}.
2: **Parameter:** Length of sliding window N_{swin} and sliding step N_{sstp}, Total number of observation points L, Total number of data in time domain T, Number of considered observation points N_{cpt} and windows N_{cwin}, Threshold α and β.
3: **Step 1: Signal compensation of X and X_{old}**
4: Conduct the division of X and X_{old} with the corresponding position of $P_R(l)$, respectively.
5: **Step 2: Compute two features**
6: **for** i in $\{1, \ldots, L\}$
7:   **for** j in $\{1, \ldots, \frac{T}{N_{\text{swin}}}\}$
8:     Conduct **compute peak** for $D = X[i, j * N_{\text{sstp}} : j * N_{\text{sstp}} + N_{\text{swin}}]$ and get $F_{\text{peak}}[i, j]$
9:     **for** k in $\{1, \ldots, N_{\text{swin}} - \alpha\}$
10:       Conduct division of $\int (D[k+\alpha] - D[k])^2$ with X_{old} and get $F_{\text{energy}}[i, j]$
11:     **end for**
12:     Convergence checking: if $F_{\text{peak}}[i, j] > \beta$, set $F_{\text{energy}}[i, j] \leftarrow 1$; otherwise, continue.
13:   **end for**
14: **end for**
15: Conduct division of $\sum (D[k + \alpha] - D[k])^2$ with X_{old} and get $F_{\text{energy}}[i, j]$
16: Conduct slicing of the F_{energy} and F_{peak} to get M_{energy} and M_{peak} with the number of m and the shape of $(N_{\text{cpt}}, N_{\text{cwim}})$
17: **Step 3: Standardization**
18: **for** M in $\{M_{\text{peak}}^{(1)}, \ldots M_{\text{peak}}^{(m)}, M_{\text{energy}}^{(1)}, \ldots M_{\text{energy}}^{(m)}\}$
19:   $M \leftarrow \frac{M - \bar{M}}{\sigma_M}$, where $\bar{M}$ is the average of M and σ_M is the standard deviation of M.
20: **Output:** Matrix of peak features $\{M_{\text{peak}}^{(1)}, \ldots, M_{\text{peak}}^{(m)}\}$ and energy features $\{M_{\text{energy}}^{(1)}, \ldots, M_{\text{energy}}^{(m)}\}$.

Additionally, we set the initial learning rate to 0.0015 and specify it to decay to 95% of the current value after every ten epochs. The batch size is 128 with 100 epochs.

3) We freeze the parameters of the two flatten layers in Table II, and the feature map concatenating from the two branches is input to LightGBM with the parameters shown in Table II. We then retrain from obtaining the optimal LightGBM model with four outputs.

We used a desktop computer with an Intel Core i7-8700 CPU running at 3.2 GHz, a GTX1080ti GPU, and 32 GB of RAM for training and verification. Considering the industrial deployment, we used TensorFlow 2.0.0 to build the model under Python 3.7.

V. CASE STUDY

In this section, we describe the evaluation of our algorithm against other methods in terms of identification and localization performance, time cost, and model size. Further, we present and discuss test results for several pipelines in service.

A. Evaluation Indices

To measure the performance of our model, we used four metrics: accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under the curve (AUC). In detail, accuracy is the ratio of the number of correctly classified categories to the total number of categories. It is a widely used index that does not consider class imbalance. Sensitivity, which is also called the true-positive rate, recall, or probability of detection in some fields, measures the proportion of positives that are correctly identified. Specificity, which is also called the true-negative rate, measures the proportion of negatives that are correctly identified.

Moreover, the AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. It is the area under the receiver operating characteristic (ROC) curve [38]. Also, unlike accuracy, the AUC is a widely used index that considers class imbalance. Specifically, for a predictor f, an unbiased estimator of its AUC can be expressed by the following *Wilcoxon-Mann–Whitney* statistic [39]

$$\text{AUC}(f) = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|} \quad (7)$$

where $\mathbf{1}[f(t_0) < f(t_1)]$ denotes an indicator function that returns $\mathbf{1}$ if $f(t_0) < f(t_1)$; otherwise, it returns $\mathbf{0}$; $\mathcal{D}^0$ is
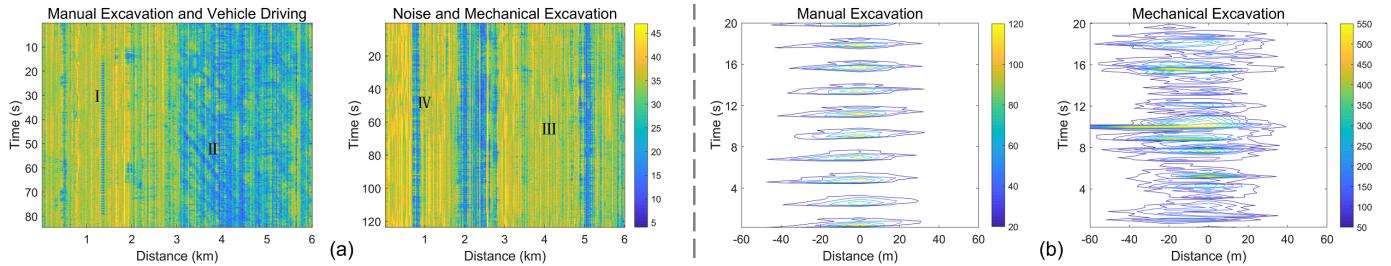
Fig. 5. Visualization results of proposed two features in Case 1. (a) Heat maps of peak features at 500 Hz. (b) Contour maps of energy features at 500 Hz.

TABLE III

RESULTS OF DIFFERENT METHODS APPLIED TO THE CASE 1 TESTING SETS (*INDICATES THE PROPOSED METHOD,
BOLD AND UNDERLINED VALUES ARE THE BEST RESULTS)

Data	Model	Background noise				Manual excavation				Mechanical excavation				Vehicle driving				Total
		Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)
100 Hz	CART_RF [40]	86.01	86.33	85.68	0.862	64.59	66.67	62.50	0.667	63.14	62.12	64.16	0.632	78.62	78.62	76.62	0.786	72.93
	LGB	87.92	88.21	87.63	0.887	72.66	70.96	74.36	0.741	72.42	71.26	73.58	0.735	83.12	90.01	76.23	0.834	78.93
	DNN [2]	86.75	86.29	87.20	0.883	90.01	89.38	90.63	0.910	89.49	90.47	88.51	0.906	88.12	87.28	88.96	0.886	88.61
	RBF_SVM [41]	93.42	94.16	92.68	0.943	89.35	91.92	86.78	0.908	85.52	86.72	84.31	0.867	89.62	88.98	90.25	0.909	89.36
	MFCNN	99.86	99.86	99.86	0.999	95.55	**96.50**	94.59	0.965	94.50	96.64	**92.36**	0.951	89.79	**97.42**	82.15	0.921	94.71
	MFCNN_LGB*	**100.0**	**100.0**	**100.0**	**1.00**	**97.23**	96.20	**98.25**	**0.977**	**95.46**	**98.64**	92.28	**0.963**	**94.15**	91.98	**96.32**	**0.938**	**96.58**
500 Hz	CART_RF [40]	88.14	86.27	90.01	0.885	77.76	75.36	80.16	0.782	78.42	74.62	82.21	0.796	83.19	85.20	81.17	0.835	81.79
	LGB	89.21	89.26	89.15	0.892	82.48	85.69	79.27	0.831	82.66	82.66	78.97	0.813	86.54	85.77	87.30	0.872	84.69
	DNN [2]	89.13	89.13	89.13	0.907	93.83	92.88	94.78	0.941	90.98	88.23	93.72	0.918	89.05	89.97	88.13	0.898	90.72
	RBF_SVM [41]	95.41	95.41	95.41	0.956	94.01	94.75	93.27	0.947	92.05	90.87	93.23	0.927	92.88	92.73	93.03	0.931	93.52
	MFCNN	**100.0**	**100.0**	**100.0**	**1.00**	99.42	**100.0**	98.83	0.996	94.45	90.62	98.27	0.951	96.34	**100.0**	92.68	0.967	97.43
	MFCNN_LGB*	**100.0**	**100.0**	**100.0**	**1.00**	**100.0**	**100.0**	**100.0**	**1.00**	**98.62**	**98.82**	**98.41**	**0.989**	**97.21**	96.55	**97.86**	**0.975**	**98.89**

the set of negative examples, and $\mathcal{D}^1$ is the set of positive examples.

B. Case 1: A 48-km Gas Pipeline of the West-East Natural Gas Transmission Project Suzhou Section

1) Data Information: We evaluated our proposed algorithm in an operational gas pipeline of the West-East Natural Gas Transmission Project Suzhou section. In detail, our evaluation was conducted at a CNPC pipeline from 10 May to 2 June and from 19 November to 17 December in 2016. The total collected data was approximately 494 GB at 500 Hz and 100 Hz, and the pipeline was approximately 48 km long with 2400 observation points, each 20 m apart, i.e., the spatial resolution was 20 m, which was sufficient in real industrial applications. Two SRAs were installed at 25 and 46 km along the pipeline. Moreover, the environment along the pipeline was complex, with farmland, factories, frequent human activity, and frequent construction, so the gathered data contained several types of noise, a low SNR, and attenuation specific to long-distance pipelines.

2) Feature Extraction Test: We applied the feature extractor introduced above and Algorithm 1 to extract the two signal features. Based on several repetitive experiments using the half-interval search method, we subjectively obtained the optimal values of the following parameters. The length of the sliding window N_{swin} and sliding step N_{sstp} were 100, the number of considered observation points N_{cpt} was 7, i.e., 120 m, and the number of considered windows N_{cwin} was 100, i.e., 20 s for 500 Hz and 100 s for 100 Hz signals. The threshold α was 2 and $\beta = 27$ and 5 for 500- and 100-Hz data, respectively. Fig. 5 shows the visualization results of the two proposed features at 500 Hz as examples, which are similar to those in the 100-Hz data. In Fig. 5(a), background noise (III) appeared

to be unchanged, whereas manual excavation (I) was clearly cyclical and caused background noise between two events of equal length. Similar to manual excavation, mechanical excavations (IV) were also periodic, but there was no obvious brief background noise between the two mechanical excavations, which had a marked energy decay in the time domain, mainly because of them having stronger transient energy than that of manual excavation. Vehicle driving (II) caused some blue-black slashes as shown in Fig. 5(a) because of the road intersecting with the pipeline; dark slashes represent cars driving on the road. The steeper the slash, the slower the vehicle; the denser the slash, the greater the traffic flow. On the contrary, similar to the peak features, the energy characteristics are periodic in Fig. 5(b). More specifically, the manual excavation period was approximately 2.2 s, whereas that of the mechanical excavation was at least 9 s, which corresponds exactly to the cycle of the event. As for the maximum values of the dimensionless energy features, those of manual excavation from 20 to 120 were much lower than those of mechanical excavation from 50 to 550.

3) Identification Performance Test: We considered the performance of the event recognizer and compared it with the other methods in the testing set. The results for an average of ten repeated experiments based on the above evaluation indices are presented in Table III for 100- and 500-Hz data. It is notable that because the datasets and the goals achieved are different, it is impossible to make a direct comparison with the baselines of previous related work; however, we tried some other methods based on the results of extracted fusion features for comparison, e.g., classification and regression tree-based random forest (CART_RF) [40], radial basis function kernel support vector machine (RBF_SVM) [41], simple deep neural network (DNN) [2], simple LightGBM, and simple MFCNN. From Table III, it can be seen that MFCNN_LGB provided
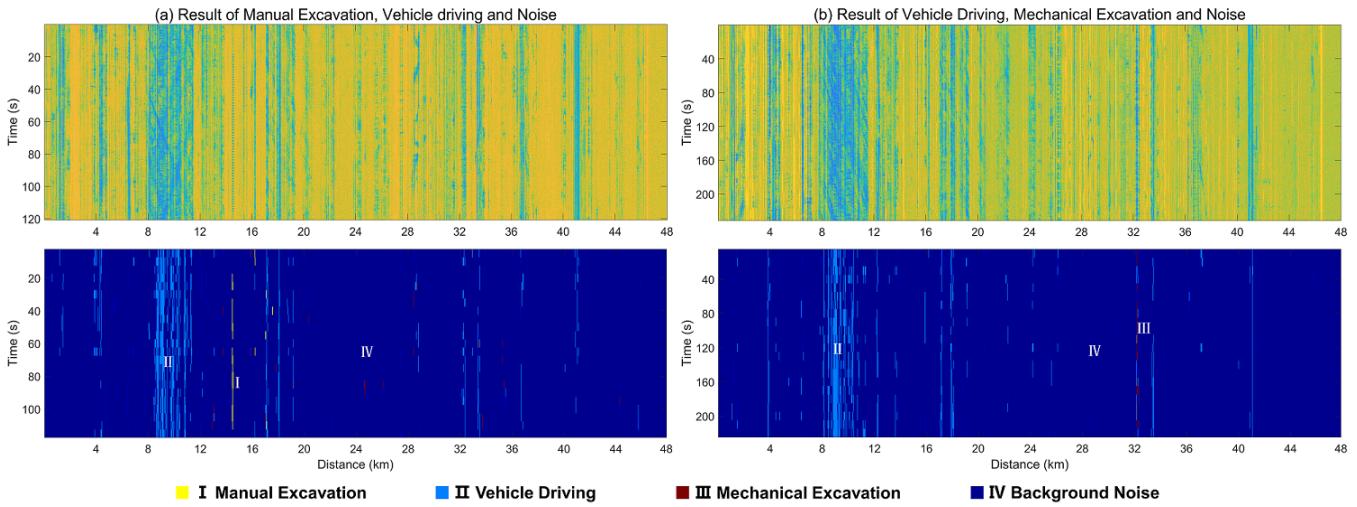
Fig. 6. Feature maps and identification results from the Case 1 pipeline. The top portion shows the peak features extracted from the feature generator and the bottom portion presents the corresponding identification results. (a) Manual excavations appear at approximately 14 km and last for nearly 110 s (result of manual excavation, vehicle driving, and noise). (b) Mechanical excavations appear at 32 km and last for approximately 210 s (result of vehicle driving, manual excavation, and noise). There are also continuous vehicle-driving events from 8 to 11 km.
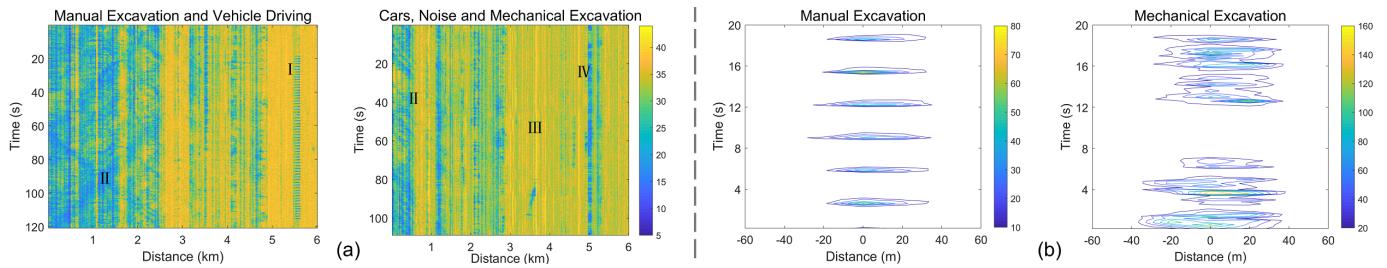

Fig. 7. Visualization results of proposed two features in Case 2. (a) Heat maps of peak features at 500 Hz. (b) Contour maps of energy features at 500 Hz.

better results than the baselines in terms of the different evaluation indices, which stems from it better integrating the two proposed high- and low-frequency features, and making the full frequency-domain features available to the model. In addition, it had better performance in all four events than the DNN and RBF_SVM, which proves that the convolution can provide more effective feature extraction for distributed signals with front-to-back frame dependency and a constant correlation length. As for simple LightGBM and CART_RF, because the input dimension of the data was large and unstructured and had local features, most of the results were still not as good as those of the other methods, despite their relatively low memory requirements and high speed. Besides, LightGBM could further fit the features obtained from MFCNN and obtain better recognition results than the fully connected layer.

4) Complete Pipeline Test: Fig. 6 shows the complete results of identification performance in Case 1 using 500-Hz data. Fig. 6(a) shows that the accuracy of our model in spatiotemporal localization and identification of manual excavation (I) was 98.63%. There were false alarms for mechanical excavation (III) at a rate of approximately 2.21%, but such samples were almost discrete and we could constrain the minimum time or observation point of the intrusion events to filter them. Fig. 6(b) shows that the model could adequately

locate and identify mechanical excavations (III) with 97.56% accuracy in the temporospatial domain.

C. Case 2: An 85-km Oil Pipeline of the National Pipeline Network Northern Pipeline Company Cangzhou Section

1) Data Information: Data were used from another real test site containing the Tianjin Port to Huabei oil pipeline of the National Pipeline Network Northern Pipeline Company Cangzhou section. The tested section was from Qingxian to Renqiu and was about 85 km in length. Along the pipeline, there were 11 large- and medium-sized river crossings, 7 ordinary railways, 13 kinds of highways, and several cities. Therefore, it was a different noise environment compared to Case 1. In addition, we collected more than 1.69 TB of data from 8 to 29 November 2020 at 100 and 500 Hz, including four complete construction days of the CNPC. Two SRAs were installed at distances of 25 and 56 km along the pipeline. The resolution in the time and space domain was the same as for Case 1.

2) Feature Extraction Test: We used the same parameters as Case 1 for feature extraction, and the results are shown in Fig. 7. The peak results shown in Fig. 7(a) are basically consistent with the results discussed for Case 1. However, there were some differences in values and periods that were

TABLE IV

RESULTS OF DIFFERENT METHODS APPLIED TO THE CASE 2 TESTING SETS (*INDICATES THE PROPOSED METHOD, BOLD AND UNDERLINED VALUES ARE THE BEST RESULTS)

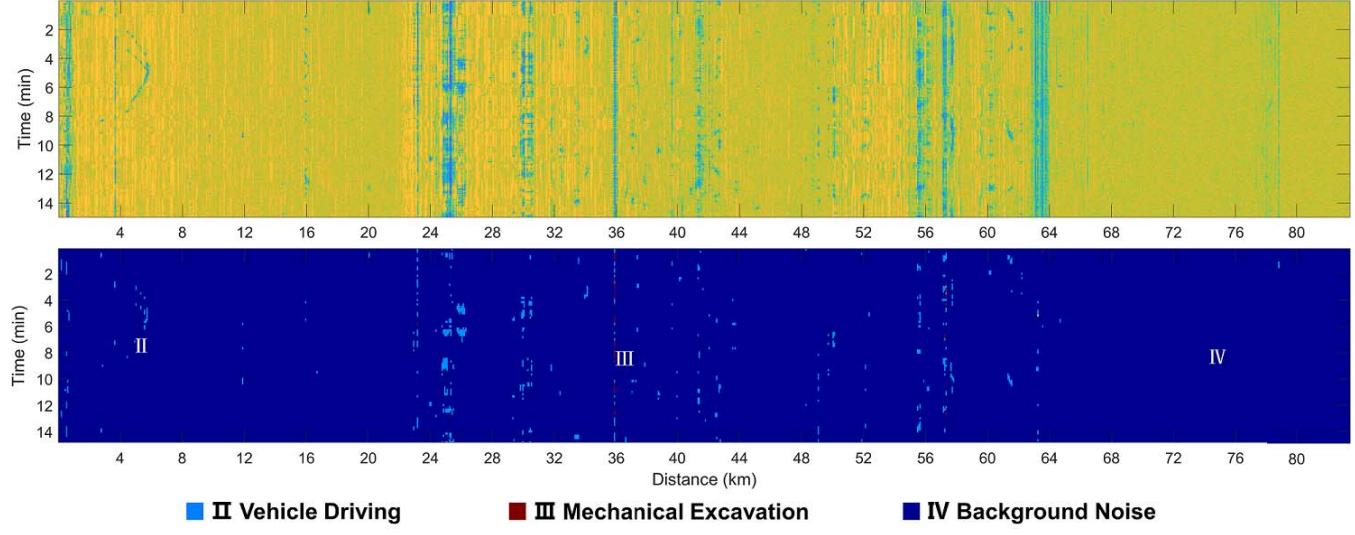| Data	Model	Background noise				Manual excavation				Mechanical excavation				Vehicle driving				Total
		Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)	Sen (%)	Spe (%)	AUC (#)	Acc (%)
100 Hz	CART_RF [40]	81.83	81.45	82.21	0.823	69.58	68.83	70.32	0.699	66.82	63.73	69.90	0.674	68.93	67.67	70.18	0.695	71.50
	LGB	82.66	81.98	83.33	0.829	73.31	72.78	73.83	0.736	70.53	71.07	69.98	0.711	78.14	77.78	78.50	0.786	76.04
	DNN [2]	87.65	86.74	88.55	0.880	86.68	86.25	87.10	0.872	82.94	83.60	82.27	0.836	81.00	79.97	82.02	0.814	84.40
	RBF_SVM [41]	93.08	91.88	94.27	0.934	86.65	85.95	87.35	0.873	85.01	85.15	84.87	0.855	82.79	84.41	81.17	0.834	86.66
	MFCNN	**98.43**	**99.14**	97.72	**0.987**	94.67	**94.01**	95.32	0.949	92.65	93.39	91.91	0.933	89.95	89.87	90.03	0.907	93.74
	MFCNN_LGB*	98.33	97.88	**98.78**	0.986	**95.28**	93.57	**96.98**	**0.958**	**95.42**	**95.64**	**95.19**	**0.960**	**91.98**	**93.14**	90.82	**0.923**	**95.13**
500 Hz	CART_RF [40]	83.73	81.79	85.67	0.845	74.44	74.30	74.57	0.753	73.21	72.54	73.87	0.736	85.26	83.25	87.27	0.862	79.18
	LGB	84.51	86.81	82.21	0.849	78.49	80.71	76.26	0.787	75.49	74.77	76.21	0.757	83.65	85.08	82.21	0.841	80.49
	DNN [2]	90.18	90.11	90.25	0.907	89.49	90.10	88.87	0.898	86.90	86.67	87.13	0.873	88.01	89.10	86.91	0.883	88.58
	RBF_SVM [41]	93.33	93.33	93.33	0.947	91.50	92.76	90.23	0.918	89.91	89.89	89.93	0.903	90.64	88.17	93.10	0.911	91.28
	MFCNN	98.44	**99.14**	97.73	0.989	95.83	96.08	**95.58**	0.961	94.49	93.93	95.05	0.948	92.74	**93.41**	92.07	0.933	95.25
	MFCNN_LGB*	**98.82**	98.62	**99.01**	**0.991**	**96.58**	**97.91**	95.25	**0.970**	**95.94**	**96.20**	**95.67**	**0.965**	**93.73**	92.78	**94.68**	**0.943**	**96.16**

Fig. 8. Feature maps and identification results from the Case 2 pipeline. The upper portion shows the peak features extracted from the feature generator and the lower portion represents the corresponding identification results. It can be clearly seen that mechanical excavation occurred for 15 min at around 36 km. Mechanical excavation was usually accompanied with signals of vehicle driving, i.e., the excavator's engine was running or the position was adjusted in order to find a better angle for excavation.

related to the burial depth of the pipeline, the relative positions of the fiber cable and pipeline, and the soil properties. The average distance from the test pipeline to the surface in Case 1 was about 1.5 m, the local climate was humid, and the soil moisture content was relatively high, which were conducive to the transmission of vibration signals. However, the average distance from the pipe to the ground surface in Case 2 was about 2.0 m and the experiment was conducted in winter, so the soil was relatively dry, which had an inhibiting effect on the transmission of vibration signals. On the contrary, the energy results shown in Fig. 7(b) were also generally consistent with those of Case 1; however, for the maximum energy values, the new data had much smaller values than those of the same intrusion event in Case 1. In addition to the reasons given for the peak results, there were also hardware deployment issues. Specifically, the 85-km pipe in Case 2 was almost twice as long as the 48-km pipe in Case 1, yet the number of SRAs used in each was the same, which resulted in a greater SNR in Case 1 than in Case 2 at the same location. However, despite the above problems, our feature extractor could still extract correct and distinct features, which indicated that our dimensionless features had good robustness.

3) Identification Performance Test: Comparing the results in Table IV with those of Case 1, our evaluation matrix values mostly had slight declines. This was caused by a decrease in

TABLE V

DEPLOYMENT PERFORMANCE TESTING WITH THE CASE 1 PIPELINE

Data	Processor	Feature time	Model time	Model size
100 Hz	CPU	4.268 s	5.268 s	
	GPU	2.956 s	3.186 s	13.79 MB
500 Hz	CPU	16.20 s	5.437 s	
	GPU	11.96 s	3.685 s	

the quality of the extracted features for the reasons referred to the information in the Case 2 feature extraction test section. Nevertheless, our model was still better than the baselines in terms of most evaluation metrics and met the requirement of real field usage.

4) Complete Pipeline Test: The complete 85-km pipeline peak features and the corresponding identification results are shown in Fig. 8. Because the environment along the pipeline was more complex in Case 2 than Case 1, it was obvious that there were more events. For example, in the range of 0−1 km, the pipe passed through a city and its signal was relatively strong, whereas 24 − 26 km was the first position of the SRA, and after verification at the real site, the original signal here had a saturation problem, resulting in the features obtained from the feature extractor being complex. Neverthe-

TABLE VI
PERFORMANCE COMPARISON OF PRACTICALLY DEPLOYABLE PSEW SYSTEMS

Group	Classifier	Monitoring distance	Positioning resolution	Accuracy	Reference
FOCUS S.L., Spain	GMM	45km	5m	68.11%	[42]
National Engineering Laboratory for Transportation Safety of Oil&gas Pipelines, China	HMM+SVM	131km	10–20m	83.0%	[43]
Future Fibre Technologies Co., Australia	–	<55km	<5m	–	[44]
Fotech Solutions, UK	–	<100km	10m	–	[45]
Omnisens, Switzerland	–	<140km	–	–	[46]
Tsinghua University, China (Our method)	MFCNN_LGB	<100km	20m	>95%	–

less, the above factors had little impact on the final recognition results. It also proves that our model has good generalization ability for the features obtained by the feature extractor.

D. Other Tests and Analysis

1) Deployment Performance Test: Model size and running time are typical practical engineering problems. They are critical to deployment performance, hardware applicability, and real application effects. Therefore, we tested a 4-min dataset from Case 1 pipeline and repeated it ten times, with the averaged results summarized in Table V. Our method could accurately identify and locate damage events within an extremely short period of time running either under a GPU (GTX 1080ti) or CPU (i7-8700 3.2 GHz) with matrix optimization and extra feature threshold recognition methods. In particular, the total recognition time was 15.64 s for 500 Hz and 6.142 s for 100-Hz signals processed by the GPU, which are almost the same as those with the CPU, and fully meet the industrial-level real-time performance requirements. The recognition response time was only about 13.58 ms per km for 500-Hz data and 5.33 ms per km for 100-Hz data under the 20 m spatial resolution and the set parameters with the GPU. Furthermore, the model size was only 13.79 MB when controlling the parameters of LightGBM, such as max_depth and min_data_in_leaf, allowing it to be deployed in most embedded systems.

2) Feature Adaptability: The applicability of the features was also verified. Despite the different factors that affect signal propagation, such as the pipeline, fiber deployment, soil moisture content, and changes in the experimental time and space, we found that our feature extractor can clearly distinguish multiple types of intrusion events, such as manual excavation, mechanical excavation, and vehicle traffic (Figs. 5 and 7). It could be attributed to the dimensionlessness of the features, which eliminated the base values and made it largely independent of changes in the environmental conditions. Obviously, from the results shown in Figs. 5–8, it can be seen that the peak features can provide professional supervisors with real-time visualization results, which has high industrial application value.

3) Modeling Spatiotemporal Stability: With the above two case studies, we have demonstrated the effectiveness of our algorithm in handling temporal and spatial signal drift. Specifically, in the time domain, we used data from May and June (summer, local temperature around 30 °C) as the training and validation set in Case 1, whereas data from November and December (winter, local temperature around 0 °C) were used as the test set. From another perspective, Case 1 was tested in 2016, whereas Case 2 was tested in 2020. As for the spatial dimension, the 2016 data for Case 1 were collected in Suzhou in central China whose high soil moisture and the shallow burial depth of the pipe were suitable for the propagation of vibration signals, whereas the 2020 data for Case 2 were gathered in Cangzhou, northern China, whose dry and sparse soil and the deep burial depth of the pipe were not suitable for the propagation of vibration signals. The distance between them was about 980 km, and there were great differences in pipeline depth, soil environment, temperature, and so on. However, the same method could be successfully used and deployed in both places. So, the above experiments show the good temporal and spatial robustness of our system.

4) Comparison of Practically Deployable PSEW Systems: The performance of some practically deployable PSEW systems is summarized in Table VI, which includes the latest researches from both academia and industry. It is clear to see that our system has good recognition and localization accuracy, thanks to the data-driven and well-fitting deep-learning methods, not just the ML methods in the other systems. However, the positioning resolution of our method is the worst, although it meets the industrial requirements for localization resolution in PSEW system [43], which requires us to continue optimizing in terms of hardware sensors and recognition algorithms.

5) Further Discussion: There are definitely still some practical deployment issues that need to be considered in more detail, such as further analysis of the discrete misidentified points in Fig. 6 and the rural misidentified points in Fig. 8. In addition, for practical applications, we can use the model compression method to further reduce the parameters of the model and apply the parallelization operation to reduce deployment cost. Furthermore, later, we should pay more attention to the issue of data utilization; i.e., using more data from nonexperimental areas for feature construction and model learning. Besides, the problem of low SNR at the end of the optical fiber between two SRAs needs to be overcome when deploying longer distance pipelines, which was partly discussed in Case 2. Most of these issues have been implemented in our experiments and deployments but are beyond the scope of this article and will be presented in our future work. Finally, as our method is based on the spatiotemporal discussions of distributed signals for signal processing, feature extraction, and event recognition, it is quite

possible to use proposed methods in other types of optical fiber sensors and improve their spatiotemporal resolution and detection performance.

VI. Conclusion

In this article, we proposed an end-to-end MFCNN and LightGBM fusion model based on the novel spatiotemporal features of distributed optical fiber sensors used to monitor the safety of long-distance oil-gas pipelines, i.e., to locate and identify third-party dangerous events in real time to ensure the safe of energy transportation. The experimental results from two real long-distance pipelines indicate that the proposed scheme can identify and locate damage events with low SNRs and little signal drift in the time and space dimensions. It demonstrated high average accuracy of 95.86% for 100-Hz data and 97.53% for 500-Hz data in the testing sets. Besides, the proposed feature generator can effectively extract the complementary and dimensionless features in a short period and has a good visualization effect, which is independent of the deployment environment and has great monitoring value. More importantly, our system fully meets the industry standards in terms of model size, real-time performance, and adaptability to different deployment conditions and environments, and has been deployed in a real long-distance energy transportation pipeline system. Therefore, our scheme provides a reference for third-party damage event recognition based on a distributed optical fiber sensor PSEW system in an open environment.

In addition, we plan to explore the possibility of achieving higher spatial resolution and shorter recognition response times, making the distributed-signal early warning system to be used in other fields of industrial inspection, measurement, and monitoring, such as national border security technology, early earthquake warning, and perimeter security systems. Besides, we are interested in applying our recognition algorithms to other optical fiber sensor systems.

References

[1] J. G. Ramírez-Camacho, F. Carbone, E. Pastor, R. Bubbico, and J. Casal, "Assessing the consequences of pipeline accidents to support land-use planning," *Saf. Sci.*, vol. 97, pp. 34–42, Aug. 2017.

[2] H. Wu, Y. Qian, W. Zhang, and C. Tang, "Feature extraction and identification in distributed optical-fiber vibration sensing system for oil pipeline safety monitoring," *Photonic Sensors*, vol. 7, no. 4, pp. 305–310, Dec. 2017.

[3] J. Li *et al.*, "Pattern recognition for distributed optical fiber vibration sensing: A review," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11983–11998, May 2021.

[4] A. H. Hartog, *An Introduction to Distributed Optical Fibre Sensors*. Boca Raton, FL, USA: CRC Press, 2017.

[5] P. Lu *et al.*, "Distributed optical fiber sensing: Review and perspective," *Appl. Phys. Rev.*, vol. 6, no. 4, Sep. 2019, Art. no. 041302.

[6] T. Chang *et al.*, "A case study on fiber optic interferometric seafloor seismic and Tsunami monitoring system in south China sea," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[7] J. Tejedor *et al.*, "Toward prevention of pipeline integrity threats using a smart fiber-optic surveillance system," *J. Lightw. Technol.*, vol. 34, no. 19, pp. 4445–4453, Oct. 1, 2016.

[8] A. Lv and J. Li, "On-line monitoring system of 35 kV 3-core submarine power cable based on φ-OTDR," *Sens. Actuators A, Phys.*, vol. 273, pp. 134–139, Apr. 2018.

[9] Y. Yang, Y. Li, T. Zhang, Y. Zhou, and H. Zhang, "Early safety warnings for long-distance pipelines: A distributed optical fiber sensor machine learning approach," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 14991–14999.

[10] G.-M. Ma, C. Shi, W.-Q. Qin, Y.-B. Li, H.-Y. Zhou, and C.-R. Li, "A non-intrusive electrical discharge localization method for gas insulated line based on phase-sensitive OTDR and Michelson interferometer," *IEEE Trans. Power Del.*, vol. 34, no. 4, pp. 1324–1331, Aug. 2019.

[11] P. J. Bock, S. Majumdar, and W. J. Bock, "Internet-based distributed data acquisition system for fiber-optic sensors," *IEEE Trans. Instrum. Meas.*, vol. 56, no. 1, pp. 32–38, Feb. 2007.

[12] Y. Lu, T. Zhu, L. Chen, and X. Bao, "Distributed vibration sensor based on coherent detection of phase-OTDR," *J. Lightw. Technol.*, vol. 28, no. 22, pp. 3243–3249, Nov. 2010.

[13] F. Tanimola and D. Hill, "Distributed fibre optic sensors for pipeline protection," *J. Natural Gas Sci. Eng.*, vol. 1, nos. 4–5, pp. 134–143, Nov. 2009.

[14] W. Ting, Y. Guo-zheng, Y. Bang-hua, and S. Hong, "EEG feature extraction based on wavelet packet decomposition for brain computer interface," *Measurement*, vol. 41, no. 6, pp. 618–625, Jul. 2008.

[15] B. M. Tabi Fouda, D. Han, B. An, X. Lu, and Q. Tian, "Events detection and recognition by the fiber vibration system based on power spectrum estimation," *Adv. Mech. Eng.*, vol. 10, no. 11, 2018, Art. no. 1687814018808679.

[16] Y. Zhang, L. Zhao, Q. Tian, and J. Fan, "Optical fiber intrusion signal recognition based on improved Mel frequency cepstrum coefficient," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2018, pp. 1–9.

[17] G. Kabir, R. Sadiq, and S. Tesfamariam, "A fuzzy Bayesian belief network for safety assessment of oil and gas pipelines," *Struct. Infrastruct. Eng.*, vol. 12, no. 8, pp. 874–889, Aug. 2016.

[18] X. Guo, L. Zhang, W. Liang, and S. Haugen, "Risk identification of third-party damage on oil and gas pipelines through the Bayesian network," *J. Loss Prevention Process Industries*, vol. 54, pp. 163–178, Jul. 2018.

[19] Z. Sheng, Z. Zeng, H. Qu, and Y. Zhang, "Optical fiber intrusion signal recognition method based on TSVD-SCN," *Opt. Fiber Technol.*, vol. 48, pp. 270–277, Mar. 2019.

[20] D. Wang and M. Li, "Stochastic configuration networks: Fundamentals and algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3466–3479, Oct. 2017.

[21] H. Wu, X. Liu, Y. Xiao, and Y. Rao, "A dynamic time sequence recognition and knowledge mining method based on the hidden Markov models (HMMs) for pipeline safety monitoring with φ-OTDR," *J. Lightw. Technol.*, vol. 37, no. 19, pp. 4991–5000, Oct. 1, 2019.

[22] M. Aktas, T. Akgun, M. U. Demircin, and D. Buyukaydin, "Deep learning based multi-threat classification for phase-OTDR fiber optic distributed acoustic sensing applications," *Proc. SPIE*, vol. 10208, Apr. 2017, Art. no. 102080G.

[23] Y. Shi, Y. Wang, L. Zhao, and Z. Fan, "An event recognition method for φ-OTDR sensing system based on deep learning," *Sensors*, vol. 19, no. 15, p. 3421, Aug. 2019.

[24] J. Tejedor, J. Macias-Guarasa, H. Martins, J. Pastor-Graells, P. Corredera, and S. Martin-Lopez, "Machine learning methods for pipeline surveillance systems based on distributed acoustic sensing: A review," *Appl. Sci.*, vol. 7, no. 8, p. 841, Aug. 2017.

[25] L. Shiloh, A. Eyal, and R. Giryes, "Deep learning approach for processing fiber-optic das seismic data," in *Proc. Opt. Fiber Sensors*, 2018, p. ThE22.

[26] S. Liehr, L. A. Jäger, C. Karapanagiotis, S. Mánzenberger, and S. Kowarik, "Real-time dynamic strain sensing in optical fibers using artificial neural networks," *Opt. Exp.*, vol. 27, no. 5, pp. 7405–7425, 2019.

[27] H. Wu *et al.*, "One-dimensional CNN-based intelligent recognition of vibrations in pipeline monitoring with DAS," *J. Lightw. Technol.*, vol. 37, no. 17, pp. 4359–4366, Sep. 1, 2019.

[28] Y. Yang, J. Li, M. Tian, Y. Zhou, L. Dong, and J. X. He, "Signal analysis of distributed optic-fiber sensing used for oil and gas pipeline monitoring," in *Proc. Int. Symp. Signal Process. Syst.*, 2019, pp. 21–25.

[29] H. Wu, C. Zhao, R. Liao, Y. Chang, and M. Tang, "Performance enhancement of ROTDR using deep convolutional neural networks," in *Proc. Opt. Fiber Sensors*, 2018, p. TuE16.

[30] W. Kong, J. Yu, J. Yang, and T. Tian, "Model building and simulation for intelligent early warning of long-distance oil & gas storage and transportation pipelines based on the probabilistic neural network," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 546, no. 2, 2020, Art. no. 022009.

[31] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" 2018, *arXiv:1805.11604*. [Online]. Available: http://arxiv.org/abs/1805.11604

[32] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[34] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.

[35] C. Lyu *et al.*, "Robust intrusion events recognition methodology for distributed optical fiber sensing perimeter security system," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[36] Y. Yang, Y. Li, and H. Zhang, "Pipeline safety early warning method for distributed signal using bilinear CNN and LightGBM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4110–4114.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[39] T. Calders and S. Jaroszewicz, "Efficient AUC optimization for classification," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Berlin, Germany: Springer, 2007, pp. 42–53.

[40] X. Wang, Y. Liu, S. Liang, W. Zhang, and S. Lou, "Event identification based on random forest classifier for ϕ-OTDR fiber-optic distributed disturbance sensor," *Infr. Phys. Technol.*, vol. 97, pp. 319–325, Mar. 2019.

[41] Z. Qu, H. Feng, Z. Zeng, J. Zhuge, and S. Jin, "A SVM-based pipeline leakage detection and pre-warning system," *Measurement*, vol. 43, no. 4, pp. 513–519, May 2010.

[42] H. Martins *et al.*, "Early detection of pipeline integrity threats using a smart fiber optic surveillance system: The pit-stop project," in *Proc. 24th Int. Conf. Opt. Fibre Sensors*, vol. 9634, 2015, Art. no. 96347X.

[43] G. Li *et al.*, "Pipe line safety monitoring using distributed optical fiber vibration sensor in the China west-east gas pipeline project," in *Proc. Asia Commun. Photon. Conf. (ACP)*, 2018, pp. 1–3.

[44] *Future Fibre Technologies Co*. Accessed: 2018. [Online]. Available: https://www.fftsecurity.com/products/fft-secure-pipe/

[45] *Fotech Solutions*. Accessed: 2021. [Online]. Available: https://www.fotech.com/solutions/pipelines/

[46] *Omnisens*. Accessed: 2021. [Online]. Available: https://www.omnisens.com/pipeline.html

Yiyuan Yang was born in Shijiazhuang, Hebei, China, in 1997. He received the B.E. degree from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei, in 2019. He is currently pursuing the M.E. degree with the Department of Automation, Tsinghua University, Beijing, China.

His current research interests include time series, optical fiber early warning systems, sensor data mining, signal processing, machine learning, and deep learning.

Haifeng Zhang was born in Jiaohe, Jilin, China, in 1986. He received the B.E. degree from the School of Physics, Liaoning University, Shenyang, Liaoning, China, in 2009, and the Ph.D. degree from the China University of Petroleum, Beijing, China.

He is currently an Assistant Researcher with the Research Institute of Tsinghua University, Pearl River Delta, Guangzhou, China. His current research interests include optical fiber early warning systems, gas and liquid multiphase flow measurement, sensor data mining, and machine learning.

Yi Li received the B.E. degree from the University of Central Lancashire, Preston, U.K., in 2004, and the Ph.D. degree from the University of Manchester, Manchester, U.K., in 2008.

He is currently an Associate Professor with the Graduate School at Shenzhen, Tsinghua University, Beijing, China. His current research interests include electrical tomography, data fusion, gas and liquid multiphase flow measurement, and image reconstruction.