SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

October 2026, Vol. 69, Iss. 10, 202303:1–202303:17
https://doi.org/10.1007/s11432-025-4792-6

# A cross-modal pre-training framework with video data for improving performance and generalization of distributed acoustic sensing

Junyi DUAN[1], Jiageng CHEN[2] & Zuyuan HE[1*]

[1]*State Key Laboratory of Photonics and Communications, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*Ningbo AllianStream Photonics Technology Co., Ltd., Ningbo 315524, China*

Received 5 August 2025/Revised 11 December 2025/Accepted 5 February 2026/Published online 27 August 2026

**Abstract** Fiber-optic distributed acoustic sensing (DAS) has emerged as a critical Internet-of-Things (IoT) sensing technology with broad industrial applications. However, the two-dimensional spatial-temporal morphology of DAS signals presents analytical challenges for conventional methods. In contrast, this complex signal structure is naturally suited for deep learning approaches like our previous work, DAS masked autoencoder (DAS-MAE). Although DAS-MAE established state-of-the-art performance and generalization without labels, it is not satisfactory in frequency analysis in temporal-dominated DAS data. Moreover, the limitation of effective training data fails to address the substantial data requirements inherent to Transformer architectures in DAS-MAE. To overcome these limitations, we present an enhanced framework incorporating short-time Fourier transform (STFT) for explicit temporal-frequency feature extraction and pioneering video-to-DAS cross-modal pre-training to mitigate data constraints. This approach learns high-level representations (e.g., event classification) through label-free reconstruction tasks. Experimental results demonstrate transformative improvements: 0.1% error rate in few-shot classification (90.9% relative improvement over DAS-MAE) and 4.7% recognition error in external damage prevention applications (75.4% improvement over from-scratch training). As the first work to pioneer video-to-DAS cross-modal pre-training, available training resources are expanded by bridging computer vision and distributed sensing areas. The enhanced performance and generalization facilitate DAS deployment across diverse industrial scenarios while advancing cross-modal representation learning for industrial IoT sensing.

**Keywords** distributed acoustic sensing, representation learning, cross-domain pre-training, self-supervised deep learning, masked autoencoder

**Citation** Duan J Y, Chen J G, He Z Y. A cross-modal pre-training framework with video data for improving performance and generalization of distributed acoustic sensing. Sci China Inf Sci, 2026, 69(10): 202303, https://doi.org/10.1007/s11432-025-4792-6

## 1 Introduction

Over the last two decades, the Internet of Things (IoT) has pursued the vision of bridging the physical and digital worlds through interconnected networks. This aims to enable pervasive, intelligent automation and data-driven decision-making across diverse applications [1,2]. Realizing this vision critically depends on deploying highly capable, intelligent sensors that can continuously gather multidimensional physical signals (e.g., acoustic, vibration, strain) while performing advanced processing. Conventional electromechanical sensors face limitations for large-scale monitoring, including deployment cost, spatial discontinuity, low tolerance to harsh conditions, and power demands. In contrast, fiber-optic distributed acoustic sensing (DAS) technology emerges as a transformative solution [3,4]. Based on phase-sensitive optical time-domain reflectometry (Φ-OTDR) [5], a DAS interrogator analyzes the phase of Rayleigh backscattering light along an optical fiber. DAS equipment effectively converts the sensing fiber into tens of thousands of continuous nodes to measure dynamic strain. This enables distributed sensing with coverage over hundred kilometers, while offering advantages such as high spatial resolution, power-free operation, and adaptability to harsh environments. These benefits have driven widespread DAS deployment in IoT applications, including earthquake monitoring [6,7], oil and gas pipeline monitoring [8,9], and railway monitoring [10,11]. Nevertheless, DAS generates spatial-temporal data (often referred to as waterfall plots), presenting unique processing challenges for intelligent interpretation. Each sensing node outputs a time-series strain signal. Combining these signals across the fiber forms a two-dimensional (2D) matrix where one axis represents distance (channel) and the other represents time. This mechanism creates data fundamentally dominated by temporal dynamics, exhibiting characteristics

* Corresponding author (email: zuyuanhe@sjtu.edu.cn)

 

distinct from natural images and more analogous to acoustic streams. Consequently, conventional image processing techniques are often ineffective for learning meaningful representations of underlying physical processes. New deep learning approaches are therefore essential to learn spatial-temporal representations directly from large volumes of raw DAS data.

Current deep learning methods for waterfall plot analysis have progressed through several distinct stages of development. Initial approaches simplified the problem by reducing 2D waterfall plots to 1D temporal signals, employing conventional deep learning architectures. Key methods demonstrated in the literature include CNN leveraging both manual and deep features [12] saved 90% of processing time compared to traditional methods. 1D-ResNet combined with SVM [13], reporting a 96.2% accuracy in identifying three-type crash accidents. End-to-end CNNs [14] attains 99.3% recognition accuracy for long-distance multi-vibration signal detection (three types of vibration events). The long short-term memory (LSTM) network outperformed CNN-based methods in 1D temporal signal event recognition, showcasing the importance of sequential modeling [15]. Other methods converted 1D temporal signals into 2D matrices suitable for input to conventional 2D CNNs using sophisticated signal processing, such as Gramian angular field [16] (enhancing temporal pattern imaging), Markov transition field [17] (capturing state transition features), or short-time Fourier transform (STFT) [18] (enhancing time-frequency analysis). While these achieve competitive accuracy (over 95%) on complex classification tasks, this artificial 2D conversion from reduced 1D signals fundamentally fails to preserve the intrinsic spatial relationships of raw waterfall plots. Consequently, the simplification eliminated the possibility of cross-channel validation, resulting in performance degradation compared to native 2D data utilization. Alternative methods have treated waterfall plots as conventional 2D images. Notably, van den Ende et al. [19] developed a self-supervised U-Net for blind denoising of earthquake DAS signals, achieving robust noise suppression without requiring clean data. Yet, this image-based method fails to account for the structural differences between natural images (dominated by spatial relationships) and waterfall plots (characterized by temporal evolution patterns). Consequently, these methods often produce suboptimal spatial-temporal representations that struggle to capture the essential dynamics of the underlying phenomena. More advanced hybrid approaches have attempted to address these limitations through sequential processing pipelines that first extract 1D temporal features before performing spatial fusion [20]. Wu et al. [21] achieved superior performance of 97% recognition accuracy, outperforming all previously mentioned network designs. Similarly, Li et al. [22] demonstrated robust intrusion detection in challenging environments, maintaining an 85.6% threat detection rate under strong background noise conditions. While these methods demonstrate improved performance through explicit temporal-spatial modeling, the sequential processing paradigm disrupts symmetry between temporal and spatial information, resulting in a compromised representation quality. Our previous work, DAS masked autoencoder (hereafter referred to as DAS-MAEv1) [23], addressed these limitations by implementing the Transformer architecture [24] within a self-supervised masked autoencoder [25]. This approach enables simultaneous processing of spatial-temporal information in waterfall plots while leveraging unlabeled data, significantly outperforming semi-supervised baselines with superior generalization capabilities. Despite this progress, two critical challenges remain: First, the embedding layers in Transformers are inadequate to capture short-time frequency characteristics and lose critical spectral information. Second, the Transformer's massive data requirements clash with DAS's data scarcity. While existing pre-processed datasets [26, 27] are directly usable, they are typically fewer than 20000 samples and statistically insufficient. Conversely, raw DAS data (e.g., the PubDAS dataset [28] contains a total data volume of 90 terabytes) demand prohibitive expert cleaning. Cross-modal pre-training thus becomes essential to bypass this bottleneck.

On the other hand, recent advances in speech processing have demonstrated remarkable success in applying Transformer architectures to spectrograms to learn temporal-frequency representation [29, 30]. From this perspective, we propose an enhanced DAS masked autoencoder (hereafter referred to as DAS-MAEv2) for waterfall plot representation learning that effectively addresses the above two key challenges. This approach begins by transforming each spatial channel's time-series signal into a 2D spectrogram using STFT, which converts an original 2D waterfall plot into a 3D spatial-temporal-frequency tensor (i.e., 3D waterfall plot) through spatial stacking of these spectrograms. Since the structural similarity between 3D waterfall plot and 3D video data, we incorporate video pre-training to simultaneously address the Transformer's data requirements while exploring domain adaptation possibilities. The DAS-MAEv2 architecture implements an asymmetric autoencoder design, where both encoder and decoder components are video vision transformers (ViViT) [31], adapted for waterfall plot analysis. During pre-training, the model is optimized through a self-supervised reconstruction task, recovering 3D waterfall plots from inputs with 90% randomly masked portions. This pre-training is done in two stages. The pre-training is implemented on video data to establish fundamental spatial-temporal representations, and then followed by domain-specific adaptation using 3D waterfall plots. Remarkably, the pre-trained model demonstrates strong clustering capabilities, automatically grouping representations from the same event types (e.g., walking, digging) without any label supervision. Experimental validation on an open benchmark dataset [26] demonstrates the framework's effectiveness. In few-shot

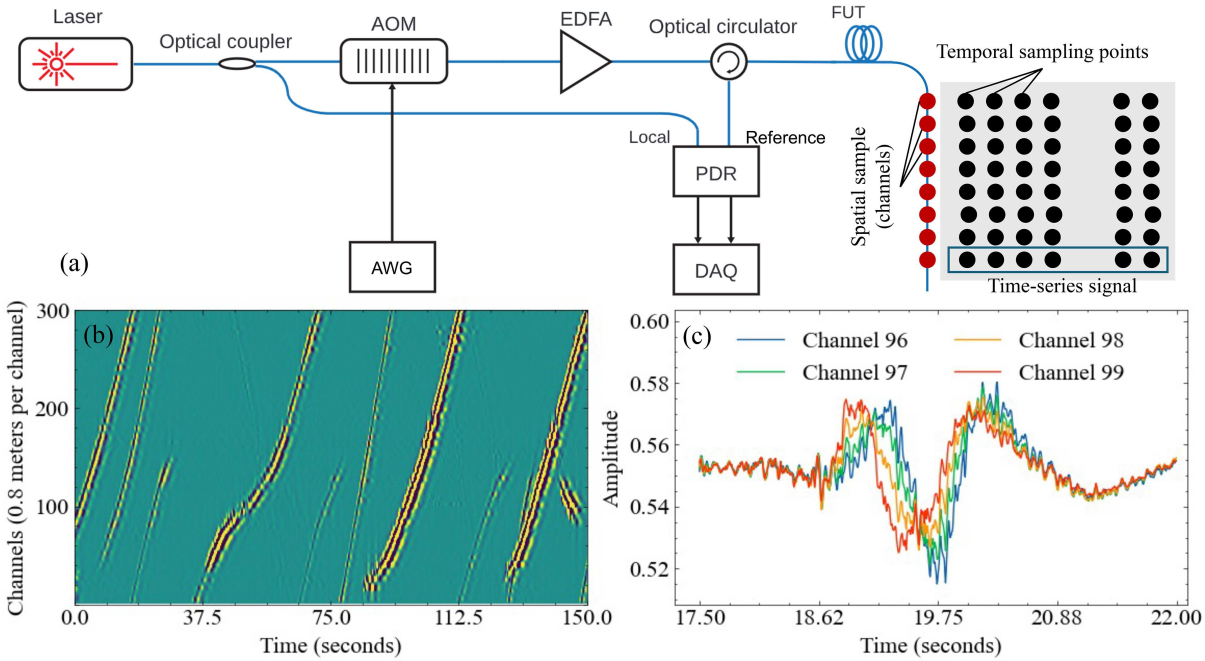


**Figure 1** (Color online) Distributed acoustic sensing system and its acquired waterfall plot. (a) Scheme of DAS. The system is composed of a laser, an acousto-optic modulator (AOM), an arbitrary waveform generator (AWG), an erbium-doped fiber amplifier (EDFA), a polarization diversity receiver (PDR), a data acquisition card (DAQ), an optical coupler, an optical circulator, and optical fibers (blue lines). A spatial segment of fiber under test (FUT) is a spatial sample or channel, which is equal to an acoustic sensor, such as a microphone. The time-series signals from each spatial sample are stacked into a 2D matrix, forming a waterfall plot. (b) A waterfall plot records the vehicle motion trajectories along a roadside. Each channel represents 0.8 m. (c) The time-series signals at adjacent channels from (b).

learning scenarios, DAS-MAEv2 achieves a 0.1% error rate (ER) (90.9% relative improvement over DAS-MAEv1) and reaches only a 5.8% error rate with only 15 labeled samples per class. Furthermore, the pre-trained model consistently outperforms from-scratch training across varying dataset sizes with over 73.9% relative improvements in practical external damage prevention applications. These results demonstrate model's high performance and strong transferability to novel applications and event types.

## 2 Methods

### 2.1 Distributed acoustic sensing and waterfall plots

Figure 1(a) illustrates the working principle of a DAS system. A coherent continuous-wave (CW) laser source emits light, which is split into two branches via an optical coupler. In the interrogation branch (upper branch), an acousto-optic modulator (AOM), driven by modulation signals from an arbitrary waveform generator (AWG), generates linear frequency-modulated (LFM) laser pulses. These pulses are amplified by an erbium-doped fiber amplifier (EDFA) and then directed through an optical circulator into the sensing fiber or the fiber under test (FUT). Rayleigh backscattering (RBS) light from the FUT propagates back through the circulator to the signal port of a polarization diversity receiver (PDR) [32]. The local branch (lower path) provides the local input for coherent detection at the PDR's local port. A data acquisition card (DAQ) samples the heterodyne signals from the PDR, followed by digital signal processing (DSP) on a computer. The external strain $\epsilon$, applied to the FUT, varies the fiber refractive index. Consequently, it introduces the RBS phase difference as

$$\Delta\Phi_L(\epsilon) = 2\beta(n + C_\epsilon)\epsilon L = \frac{\epsilon L}{K_\Phi}, \tag{1}$$

where $\beta = 2\pi/\lambda$ is the wave vector in vacuum, $n$ is the fiber core refractive index, $C_\epsilon$ is the photo-elastic coefficient, $L$ is the spatial differential distance, and $K_\Phi$ is sensitivity coefficient. For a laser wavelength of 1550 nm and a fiber core refractive index $n = 1.46$, the sensitivity coefficient $K_\Phi$ is around 110.37 $n\epsilon$·m/rad [33,34]. Eq. (1) establishes a linear relationship between the strain (vibration waveform) and the RBS phase difference. Crucially, the phase difference remains zero outside vibration-active regions, enabling precise vibration localization. The hundred-kilometer-long FUT is transformed into tens of thousands of sensing nodes (spatial samples) by DAS. Each sensing

node measures the strain and outputs a corresponding time-series signal. By combining the time-series signals from multiple sensing nodes, a 2D spatial-temporal matrix is formed as a waterfall plot.

Figures 1(b) and (c) illustrate the vehicle motion trajectories recorded by DAS along a roadside, manifesting as non-stationary spatial-temporal signals across adjacent channels and sampling points. Unlike 2D images, the horizontal axis of waterfall plots represents time, while the vertical axis corresponds to spatial position along the sensing fiber (FUT). Moreover, the temporal sample rate is significantly higher than the spatial sample rate in DAS measurements. For instance, the temporal sample rate is 200 samples per second, compared to the spatial sample rate of 1.25 samples per meter (i.e., 0.8 m per channel) in Figure 1. This results in an asymmetric spatial-temporal information density, where temporal information dominates in information entropy. Furthermore, the signal amplitude in DAS measurements depends on both vibration intensity and fiber-medium coupling efficiency, while the coupling efficiency varies significantly along the sensing fiber in practice. As shown in Figure 1(b) (signals in 37.5 to 75.0 second interval), channels before 200 exhibit stronger fiber-ground coupling than subsequent channels, causing amplitude variations while preserving spectral pattern similarity. This physical constraint necessitates prioritized analysis of frequency-domain features when processing time-series signals in waterfall plots. Figure 1(c) displays the time-series signals from channels 96 to 99. These signals contain rich waveform and frequency information, and adjacent channels exhibit substantial redundancy (similar waveforms or waveform coherence [35]). Prior studies [27,36] demonstrated that focusing on temporal feature extraction within waterfall plots substantially improves learned representations. However, exclusive reliance on temporal information, while neglecting spatial correlations, could yield suboptimal results. This distinctive characteristic of waterfall plots establishes a unique spatial-temporal coupling, fundamentally different from that observed in natural images.

### 2.2 DAS-MAEv2

Figure 2(a) illustrates the DAS-MAEv2 framework for self-supervised masked autoencoding of waterfall plots. A DAS waterfall plot is represented as $\mathbf{x} \in \mathbb{R}^{C\times S}$, where $C$ denotes the number of spatial channels and $S$ represents temporal sampling points. For each spatial channel $i = 0, 1, \cdots, C-1$, we compute the STFT transformation of $\mathbf{x}[i]$ by

$$\begin{aligned}
&\mathbf{X}[i,t,f] = \left| \sum_{n=0}^{L-1} \mathbf{x}_{\text{pad}}[i, n+t\cdot P]\cdot w[n]\cdot e^{-j2\pi fn/Q} \right|,\\
&\mathbf{x}_{\text{pad}}[i, n+t\cdot P] = \begin{cases} \mathbf{x}[i, n+t\cdot(P-M)], & 0 \leqslant n < L-M,\\ 0, & L-M \leqslant n < L, \end{cases}\\
&i = 0,1,\cdots,C-1, \quad t = 0,1,\cdots,T-1, \quad f = 0,1,\cdots,Q-1.
\end{aligned} \tag{2}$$

Here, $i$ indexes the spatial channel, $t$ indexes the time frame, $f$ indexes the frequency bin, and $n$ indexes the sampling point. The function $|\cdot|$ denotes the absolute value, $w[n]$ is a Hamming window with length of $L$, $P$ is the hop length, and $Q$ is the FFT size. In our implementation, we set $L = P$ to ensure no overlap between adjacent temporal segments, which prevents redundant information and potential information leakage during the masked reconstruction in DAS-MAEv2. Additionally, matching the window length with the FFT size ($L = Q$) adheres to signal processing conventions and preserves energy conservation according to Parseval's theorem. The time-domain zero-padding ($M$ zeros) interpolates the frequency spectrum to provide better frequency bin localization. The STFT transforms the input waterfall plot $\mathbf{x}$ into a spatial-temporal-frequency data $\mathbf{X} \in \mathbb{R}^{C\times T\times Q}$, where $T = \lfloor S/L \rfloor$ is the number of time frames, and $\lfloor\cdot\rfloor$ is the floor function. Due to the Hermitian symmetry of the FFT, only the first $F = Q/2$ frequency bins are retained for model training, resulting in a reduced tensor $\mathbf{X} \in \mathbb{R}^{C\times T\times F}$, i.e., 3D waterfall plot. This tensor is then partitioned into $N$ non-overlapping spatial-temporal-frequency tubes $\{\mathbf{X}_i \in \mathbb{R}^{C_p\times T_p\times F_p}\}_{i=1}^{N}$, where $N = \lfloor C/C_p \rfloor \times \lfloor T/T_p \rfloor \times \lfloor F/F_p \rfloor$. $C_p$, $T_p$, and $F_p$ denote the number of channels, time frames, and frequency bins in a tube, respectively. During pre-training, we implement aggressive random masking by removing 90% of tubes ($N_m = \lfloor 0.9N \rfloor$) following a uniform distribution (referred to as "random sampling"). The remaining $N_v = N - N_m$ visible tubes $\mathbf{X}_{\mathbf{m}^c}$ (where $\mathbf{m}^\mathbf{c}$ denotes the complement of the mask set $\mathbf{m}$) are encoded by the DAS-MAEv2 encoder into latent representations. A lightweight decoder estimates the masked tubes $\hat{\mathbf{X}}_\mathbf{m}$ from representations and mask tokens. The reconstruction objective minimizes the normalized mean square error (MSE) across masked regions within the entire dataset $\mathcal{X}$:

$$L_{rec} = \mathbb{E}_{\mathbf{X}\sim\mathcal{X}} \left[ \frac{1}{N_m} \sum_{i=1}^{N_m} \left\| \mathbf{X}_\mathbf{m}^{(i)} - \hat{\mathbf{X}}_\mathbf{m}^{(i)} \right\|_2^2 \right]. \tag{3}$$

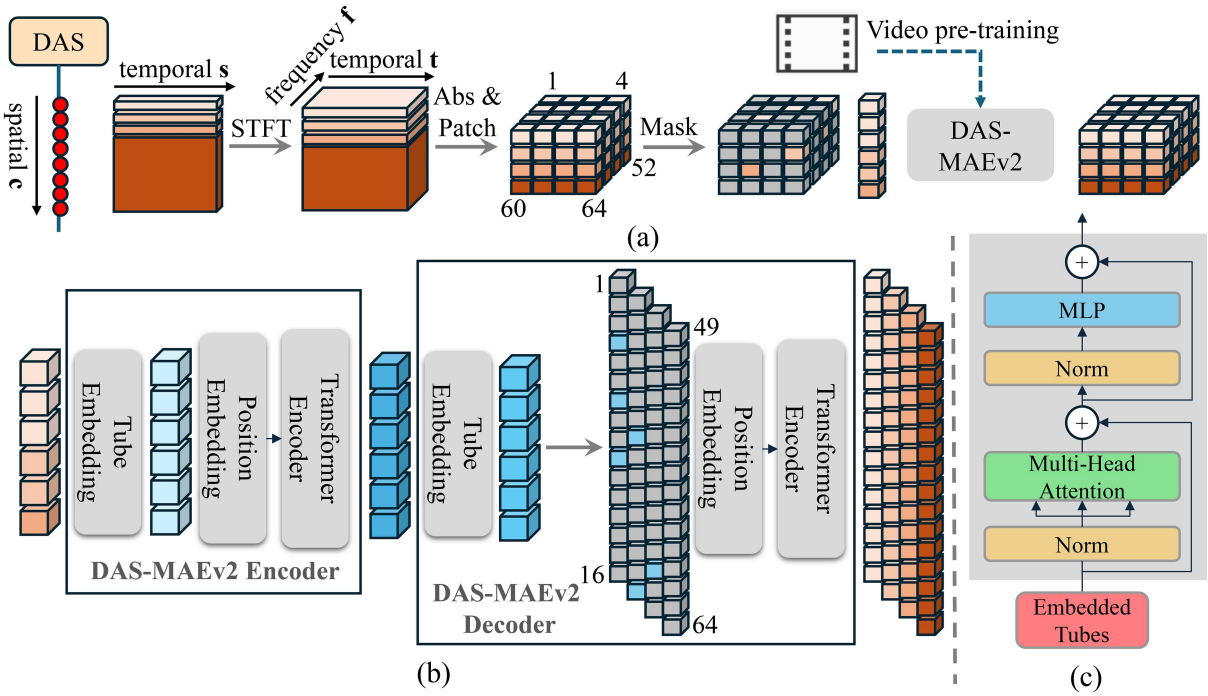


**Figure 2** (Color online) Architecture of DAS-MAEv2 in pre-training. (a) The framework of DAS-MAEv2. Its hyperparameters are given in Table 1. (b) The inside modules of DAS-MAEv2. It employs the video vision transformers (ViViT) [31] for both the encoder and decoder, each including a tube embedding layer, a position embedding layer, and a Transformer encoder. It is important to note that the encoder is deliberately designed with a larger size compared to the decoder (see Table 1). (c) The structure of a Transformer encoder (depth of $L$, head of $H$) [37]. It consists of $L$ Transformer blocks, each including multi-head attention ($H$ heads), an MLP block, and norm layers.

The 90% mask ratio (the ratio of removed tubes) largely eliminates the waterfall plot's redundancy and prevents interpolation solutions. The lightweight decoder design forces the encoder to generate high-level representations from data points.

As illustrated in Figure 2(b), the DAS-MAEv2 framework employs a ViViT [31] for both its encoder and decoder. The encoder begins by processing visible input tubes $\mathbf{X_{m^c}} \in \mathbb{R}^{N_v \times C_p \times T_p \times F_p \times D_i}$ (where $D_i = 1$ is an additional data dimension for network processing) through a 3D convolutional layer that projects the tube dimension $D_i$ to $D_e$. Then the tubes are flattened as $\mathbb{R}^{N_v \times C_p \times T_p \times F_p \times D_e} \rightarrow \mathbb{R}^{N_v \times (C_p \cdot T_p \cdot F_p) \times D_e}$. To preserve positional information, standard 1D learnable positional encodings [24] are incorporated before the sequence of embedded tubes undergoes processing through $L_e$ stacked Transformer blocks. Each Transformer block, demonstrated in Figure 2(c) [37], comprises multi-head attention with $H_e$ heads, MLP layers, and normalization operations. Specifically, multi-head attention weights cross-tube dependencies across parallel heads. The MLP operates independently on each tube and provides non-linear transformation capabilities. Normalization stabilizes training by standardizing activations. The encoder ultimately yields high-level representations $\mathbf{Z_{m^c}} \in \mathbb{R}^{N_v \times (C_p \cdot T_p \cdot F_p) \times D_e}$. The decoder operates on these representations $\mathbf{Z_{m^c}}$ by first applying tube embedding (a linear projection) to dimension $D_d$. Then, these representations reconstruct the complete set of $N$ tokens by inserting a shared learnable mask token $\mathbb{R}^{1 \times (C_p \cdot T_p \cdot F_p) \times D_d}$ at masked positions. The 1D position embeddings are also added to maintain spatial-temporal-frequency relationships. The sequence is then transformed through $L_d$ Transformer blocks ($H_d$ heads) before final reshaping to match the original input dimensions $\mathbb{R}^{N \times C_p \times T_p \times F_p \times D_i}$. Complete architectural specifications, including tensor dimensions and hyperparameter values are systematically presented in Table 1.

DAS-MAEv2 significantly advances beyond DAS-MAEv1 by incorporating STFT to enable joint temporal-frequency analysis. This approach addresses critical limitations in the original DAS-MAEv1 framework, where the raw temporal series are directly time-embedded via employing a DAS-ViT architecture (analogous to Vision Transformers for images). This time-embedded overemphasized temporal amplitude variations (theoretically linear to vibration intensity in ideal DAS conditions) while being limited by spatial variations in fiber-medium coupling efficiency. These variations introduce non-linear distortions between the measured amplitude and the actual vibration energy. The STFT transformation converts 2D spatial-temporal matrices into interpretable 3D spatial-temporal-frequency tensors that explicitly reveal frequency-domain patterns while maintaining temporal fidelity. This transformation provides more robust features, particularly through instantaneous frequency shifts that are

**Table 1** Pre-training framework of DAS-MAEv2.

| Module/variable | Sub-module | Layer | Hyperparameter | Output size/size | Value |
|---|---|---|---|---|---|
| $\mathbf{x}$ | – | – | – | $\mathbb{R}^{C\times S\times D_i}$ | $C_p=2$ |
| $\mathbf{X}$ (STFT & patched) | – | – | – | $\mathbb{R}^{N\times C_p\times T_p\times F_p\times D_i}$ | $T_p=16$ |
| $\mathbf{X_{mc}}$ (masked) | – | – | – | $\mathbb{R}^{N_v\times C_p\times T_p\times F_p\times D_i}$ | $F_p=16$ |
| Encoder | Tube embedding | 3D Conv | kernel = stride = $C_p\times T_p\times F_p$ | $\mathbb{R}^{N_v\times(C_p\cdot T_p\cdot F_p)\times D_e}$ | $N=216$ |
| | Position embedding | – | – | $\mathbb{R}^{N_v\times(C_p\cdot T_p\cdot F_p)\times D_e}$ | $N_v=21$ |
| | Transformer encoder | Transformer block | depth = $L_e$, head = $H_e$ | $\mathbb{R}^{N_v\times(C_p\cdot T_p\cdot F_p)\times D_e}$ | $D_e=384$ |
| $\mathbf{Z_{mc}}$ | Representations (masked) | – | – | $\mathbb{R}^{N_v\times(C_p\cdot T_p\cdot F_p)\times D_e}$ | $D_d=192$ |
| Decoder | Tube embedding | Linear | in_channel = $D_e$, out_channel = $D_d$ | $\mathbb{R}^{N_v\times(C_p\cdot T_p\cdot F_p)\times D_d}$ | $L_e=12$ |
| | Position embedding | – | – | $\mathbb{R}^{N\times(C_p\cdot T_p\cdot F_p)\times D_d}$ | $H_e=6$ |
| | Transformer encoder | Transformer block | depth = $L_d$, head = $H_d$ | $\mathbb{R}^{N\times(C_p\cdot T_p\cdot F_p)\times D_d}$ | $L_d=4$ |
| $\hat{\mathbf{X}}$ | – | – | – | $\mathbb{R}^{N\times C_p\times T_p\times F_p\times D_i}$ | $H_d=3$ |

less susceptible to coupling artifacts. Consequently, the usage of STFT transformation facilitates more effective representation learning for DAS. The STFT results, i.e., 3D waterfall plots (spatial-temporal-frequency), require replacing DAS-ViT with ViViT for 3D tensor processing. The structural similarity between 3D waterfall plots (spatial-temporal-frequency) and video data (temporal-spatial-spatial) inspires a dual-stage pre-training approach. It involves initial masked reconstruction on video data and is followed by further pre-training on 3D waterfall plots. Comprehensive ablation studies in Section 4 demonstrate the individual contributions of STFT integration, dual-stage pre-training, and mask-reconstruction hyperparameter settings to learned representation quality.

### 2.3 Dual-stage pre-training strategy

The DAS-MAEv2 framework (∼23M parameters) employs a dual-stage pre-training approach. In the first stage, the model is pre-trained on video data, where we directly use the trained parameters from prior work [38] since it is not the primary focus of our current study. The second stage involves domain-specific adaptation through continued pre-training on an open-access waterfall plot dataset contributed by Cao et al. [26]. This comprehensive dataset consists of approximately 15000 samples, with each sample containing temporal sequences of 12 channels × 10000 sampling points, acquired at sample rates of either 12.5 or 8.0 kHz. Notably, our model imposes no specific requirements on the channels or sampling points of the input signal. The fixed channel number and temporal sequence length in this dataset are merely representative of the used data. The dataset encompasses six distinct event categories: background noise, digging, knocking, watering, shaking, and walking. The training set contains about 12000 samples (approximately 2000 samples per event category), while the testing set comprises 3000 samples (roughly 500 samples per event category).

For the domain adaptation stage, we utilize only the raw data without any class labels during pre-training. According to Subsection 3.5, the proposed method requires input signals to maintain SNR $>0.1$ for reliable operation, while optimal performance is achieved when SNR $\geqslant 1$. The training configuration employs a batch size of 64 across 500 epochs, using the AdamW optimizer [39] with a cosine decay learning rate schedule (initial learning rate of $10^{-3}$, 40-epoch warmup period) [40]. The complete pre-training process requires approximately 5 h on an NVIDIA RTX 4090 GPU. The pre-trained model is able to achieve real-time inference (4 ms @ RTX 4090), which is suitable for field deployment.

### 2.4 Evaluating the pre-trained DAS-MAEv2

To assess the learned representation quality, we evaluate DAS-MAEv2's performance on a classification task with labeled dataset $\mathcal{D}=\{(\mathbf{x}^{(k)},\mathbf{y}^{(k)})\}_{k=1}^{K}$ where $\mathbf{y}\in\{0,1\}^M$ is a one-hot encoded label over $M$ event classes. Given an input $\mathbf{x}\in\mathbb{R}^{C\times S}$, the pre-trained encoder $\mathbf{E}$ generates latent representations $\mathbf{Z}\in\mathbb{R}^{N\times(C_p\cdot T_p\cdot F_p)\times D_e}$ without masking. Consequently, the representation length changes from $N_v$ to $N$. These representations are aggregated via averaging along the first dimension:

$$\bar{\mathbf{Z}}=\underset{\text{(1st dim)}}{\text{Mean}}\big(\mathbf{E}[\text{STFT}(\mathbf{x})]\big), \tag{4}$$

where STFT$(\cdot)$ is the short-time Fourier transform. The average representation $\bar{\mathbf{Z}}$ is then mapped to class probabilities $\hat{\mathbf{y}}$ via two protocols.

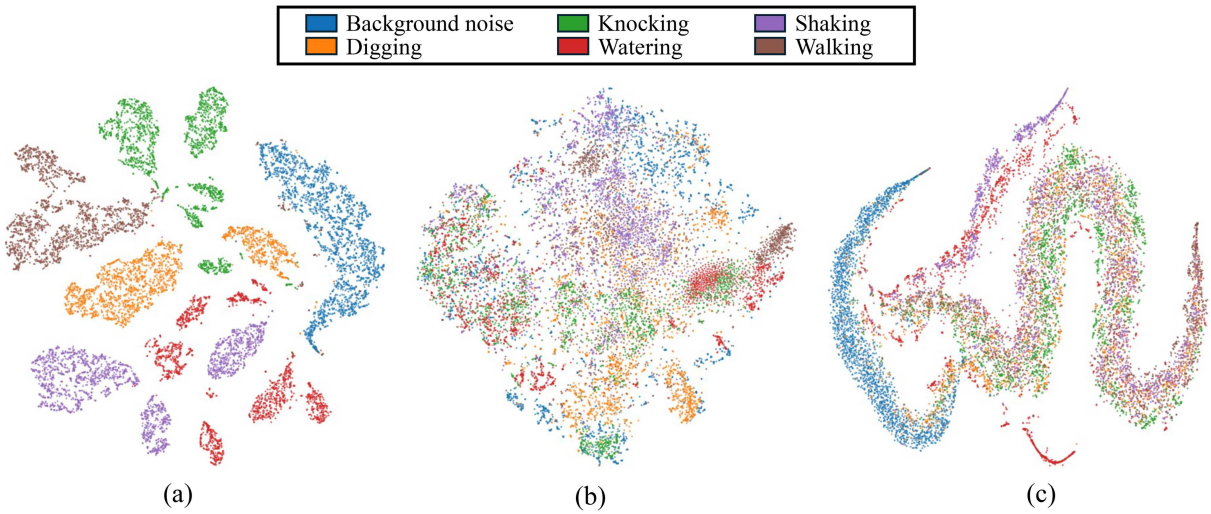


**Figure 3** (Color online) The t-SNE visualization of representations. Labels are used to assign colors and evaluate the quality of learned representations. (a) Averaged representations $\bar{\mathbf{Z}}$ from DAS-MAEv2; (b) representations learned by principal component analysis (PCA); (c) representations obtained from the STFT transformation of raw data [26].

Linear probing uses a lightweight linear classifier $\mathbf{D}$, which is a single linear layer, to project average representation $\bar{\mathbf{Z}}$ to class probabilities, while the encoder $\mathbf{E}$ remains frozen:

$$\arg\min_{\mathbf{D}} \mathbf{L}_{cls}\left[\mathbf{y}, \mathbf{D}\left(\underset{(1\text{st dim})}{\text{Mean}}\left(\mathbf{E}[\text{STFT}(\mathbf{x})]\right)\right)\right], \tag{5}$$

where $\mathbf{L}_{cls}(\cdot)$ is the cross-entropy loss [41] for classification. Linear probing evaluates whether class-relevant representations emerge as linearly separable clusters in the latent space. It directly reflects the encoder's reconstruction-driven representation quality.

Fine-tuning attempts to use task-specific nonlinear features and optimizes both the encoder $\mathbf{E}$ and classifier $\mathbf{D}$ (a single linear layer):

$$\arg\min_{\mathbf{E},\mathbf{D}} \mathbf{L}_{cls}\left[\mathbf{y}, \mathbf{D}\left(\underset{(1\text{st dim})}{\text{Mean}}\left(\mathbf{E}[\text{STFT}(\mathbf{x})]\right)\right)\right]. \tag{6}$$

This method demonstrates the model's adaptability. It leverages minimal task-specific training (far fewer epochs than supervised training from scratch) to maximize transfer performance.

## 3 Experiments

### 3.1 Visualization of learned representations

We first conducted a qualitative evaluation of DAS-MAEv2's learned representations through visualization using t-distributed stochastic neighbor embedding (t-SNE) [42]. As a nonlinear dimension reduction technique, t-SNE effectively preserves clustering patterns from high-dimensional spaces in low-dimensional projections. For our analysis, the pre-trained encoder extracted average representations $\bar{\mathbf{Z}}$ from the training dataset. The obtained $\bar{\mathbf{Z}}$ were further projected to the 2D space using t-SNE with perplexity of 40, learning rate of 2000, and 500 iterations. Class labels were used solely for color assignment in Figure 3 to enhance visual discrimination.

Figure 3(a) demonstrates that DAS-MAEv2 generates well-separated clusters for all six event categories, with minimal inter-class overlap. The background noise class (blue) forms a compact cluster, which is consistent with its stationary nature. Although dynamic events (walking, knocking, etc.) show intra-class dispersion, our ablation studies in Section 4 confirm that this can be effectively resolved through supervised fine-tuning and achieve less than 0.1% error rate. For comparative evaluation, we contrast these representations with those from principal component analysis (PCA) [43]. Figure 3(b) shows that PCA produces completely overlapping clusters and fails to discriminate between event types. This result reveals the PCA's limitation of linear projections. While Figure 3(c) shows that STFT-transformed waterfall plots provide some separable spectral representations between background noise and dynamic events (unlike PCA), these representations still cannot achieve the clear inter-class separation

**Table 2** Few-shot learning error rates (%) of different models. The best results are in bold.

| Data No. | DAS-MAE | | | ACAB [44] | RI | |
|---|---|---|---|---|---|---|
| (per class) | v1 [23] | v2-w/o-video | v2 | | (DAS-MAEv1, DAS-MAEv2) | (ACAB, DAS-MAEv2) |
| 15 | 10.0 | 7.1 | **5.8** | 16.5 | 42.0 | 64.8 |
| 40 | 3.6 | **1.1** | **1.1** | 6.2 | 69.4 | 82.3 |
| 80 | 2.8 | 0.6 | **0.3** | 4.4 | 89.3 | 93.2 |
| 120 | 1.3 | 0.4 | **0.2** | 3.8 | 84.6 | 94.7 |
| 205 | 1.1 | 0.2 | **0.1** | 3.1 | 90.9 | 96.8 |

**Table 3** Comparison of DAS-MAEv2 to other models. The best results are in bold.

| Year | Model | Backbone | Training paradigms | Error rate (%) |
|---|---|---|---|---|
| 2025 | DAS-MAEv2 (ours) | Transformer | Self-supervised | **0.06** |
| 2025 | DAS-MAEv1 [23] | Transformer | Self-supervised | 0.32 |
| 2025 | ODAS [46] | ResNet-18 | Semi-supervised & Contrastive | 0.75 |
| 2025 | DMFCC-MSPGAN [47] | CNN | GAN | 2.20 |
| 2025 | ResNet-34 [48] | ResNet-34 | Knowledge distillation | 1.69 |
| 2024 | ACAB [44, 49] | CNNs-BiLSTM | Semi-supervised | 0.10 |
| 2024 | ST-T [50] | Transformer | Supervised | 1.50 |
| 2023 | VTTCG-DD [51] | Dendrite Net | Supervised | 1.40 |

demonstrated by DAS-MAEv2 in Figure 3(a). These visual comparisons provide empirical evidence that DAS-MAEv2 learns semantically meaningful representations without label supervision.

### 3.2 Quality of learned representations

To evaluate the learned representations, we compared DAS-MAEv2 against DAS-MAEv1 and the semi-supervised ACNN-SA-BiLSTM (ACAB) model [44] on the open dataset [26]. Since the ACAB model was originally trained under the Mean Teacher framework [45] with limited labeled data, we ensured an equitable evaluation by fine-tuning all DAS-MAE versions under identical few-shot learning constraints (i.e., using the same labeled data subsets). For fairness, all models were fine-tuned using AdamW (initial learning rate = $10^{-5}$, weight decay = 0.05, 4 warm-up epochs) with a cosine scheduler for 50 epochs [39]. We quantify improvements via the relative improvement (RI):

$$\mathrm{RI}(\mathrm{A},\mathrm{B}) = \frac{\mathrm{ER_A} - \mathrm{ER_B}}{\mathrm{ER_A}} \times 100\%, \tag{7}$$

where $\mathrm{ER_A}$ and $\mathrm{ER_B}$ denote the error rates of models A and B, respectively. Here, $\mathrm{RI}(\mathrm{A},\mathrm{B})$ explicitly measures the percentage reduction in error rate when replacing model A with model B. Note that RI is asymmetric: $\mathrm{RI}(\mathrm{A},\mathrm{B}) \neq \mathrm{RI}(\mathrm{B},\mathrm{A})$, as it normalizes improvement relative to the baseline model (A in this case). We utilize these metrics to quantify the model's effectiveness in few-shot scenarios.

The experimental results presented in Table 2 demonstrate the superior classification performance of our DAS-MAEv2 model compared to both the previous DAS-MAEv1 version and the semi-supervised ACAB model. With only 15 labeled samples per class (90 total), DAS-MAEv2 achieves an ER of 5.8%, representing a significant 64.8% RI over ACAB's 16.5% ER and a 42.0% improvement over DAS-MAEv1's 10.0% ER. This performance advantage becomes even more pronounced as the number of labeled samples increases. DAS-MAEv2 reaches an exceptional 0.1% ER with 205 samples per class, corresponding to a 96.8% RI over ACAB's 3.1% ER and a 90.9% RI over DAS-MAEv1's 1.1% ER. The 'v2-w/o-video' variant, which incorporates STFT transformations but excludes video data pre-training, already shows substantial improvements over DAS-MAEv1 (e.g., 7.1% vs. 10.0% ER at 15 samples). While the full DAS-MAEv2 model with video data further achieves additional performance gains (7.1% vs. 5.8% ER at 15 samples). These results not only validate the effectiveness of our architectural improvements but also highlight DAS-MAEv2's remarkable capability to achieve practical deployment requirements ($\leqslant$10% ER) with minimal labeled data. The progressive increase in RI values with more training data (from 42.0% to 90.9% for DAS-MAEv1 to DAS-MAEv2) further demonstrates DAS-MAEv2's superior representation learning ability compared to both its predecessor and the semi-supervised ACAB model.

To further compare with other existing DAS models, DAS-MAEv2 was fine-tuned on the whole open dataset. Table 3 [23,44,46–51] presents the classification results (error rate), where our method achieves superior performance with the lowest error rate of 0.06%.

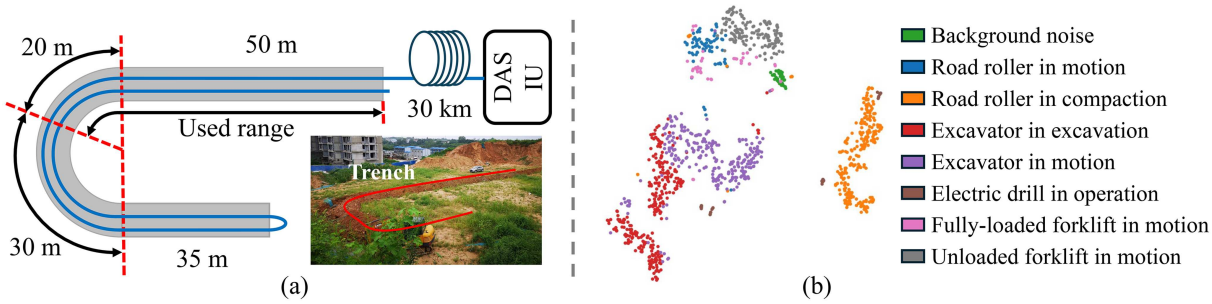


**Figure 4** (Color online) Field experiment of external damage prevention. (a) Experimental layout. The experimental configuration features a U-shaped trench containing the buried sensing fiber. The approximately 135-m-long trench maintains an average depth of 1 m, consisting of three distinct segments: a 50-m linear section, a 50-m curved section, and a 35-m linear section. The waterfall plot is only captured within the first 50 m and the 20 m of the curved section ('used range'). (b) t-SNE visualization of representations for 8 novel event types that were completely unseen during DAS-MAEv2 pre-training. Despite the lack of supervision, the proposed model successfully maps these novel events into distinct, well-separated clusters.

### 3.3 Generalization analysis on novel events

In real-world industrial IoT deployments, the environment is highly dynamic. New types of events or threats (referred to as open-set samples) can emerge unexpectedly. Therefore, a robust model must possess strong generalization capabilities to generate discriminative representations for these unforeseen events without immediate retraining. To evaluate this capability, we adopt an open-set testing protocol. Unlike standard supervised evaluation, this protocol tests the pre-trained model directly on a set of completely novel classes. The quality of generalization is assessed by analyzing the distribution of embedded representations in the latent space. Suppose the model has learned intrinsic signal structures; novel events should naturally form distinct clusters based on their physical characteristics.

To thoroughly evaluate the model generalization to novel events under real-world conditions, we conducted field experiments within an external damage prevention application deployed in Zhengzhou, China. The experimental configuration (Figure 4(a)) utilized a U-shaped fiber trench containing three geometrically distinct sections (50-m linear, 50-m curved, and 35-m linear segments), designed to replicate practical fiber deployment scenarios with both straight and curved paths. This folded fiber configuration enables sensing through differential fiber-medium coupling, providing two independent observations of each event. The vibration was monitored by a DFVS-850 DAS system operating at a sample rate of 2 kSa/s with 10-m spatial resolution. Notably, our analysis intentionally focuses on the first 70 m, i.e., 'used range', of fiber to eliminate redundancy of being detected twice at both linear sections. This prevents potential model bias from cross-validating duplicate event features across different linear sections. We established a carefully designed vibration dataset consisting of eight distinct categories.

(1) **Background noise (Class 0)**: Baseline recordings captured under static conditions with no active vibration sources, comprising 42 samples.

(2) **Road roller in motion (Class 1)**: Pure movement of the road roller along the fiber optic cable path without compaction operations, including 100 samples.

(3) **Road roller in compaction (Class 2)**: Combined movement and dynamic vertical compaction forces applied by the road roller, with 300 collected samples.

(4) **Excavator in excavation (Class 3)**: Full excavation activities were performed using two different machinery configurations, including both light-duty (6-ton) and heavy-duty (22-ton) excavators. This class contains 332 samples.

(5) **Excavator in motion (Class 4)**: Pure movement of excavators along the fiber path, captured in 334 samples.

(6) **Electric drill in operation (Class 5)**: High-frequency vibrational signals generated by handheld electric drilling equipment, represented by 38 samples.

(7) **Fully-loaded forklift in motion (Class 6)**: Movement of a 9-ton forklift operating at maximum cargo capacity (total weight 11 tons), comprising 54 samples.

(8) **Unloaded forklift in motion (Class 7)**: Movement of the same 9-ton forklift in unloaded configuration, with 186 collected samples.

The complete dataset contains approximately 1400 vibration events in total. We applied our pre-trained DAS-MAEv2 encoder to this prevention dataset, whose events were completely excluded from the model's pre-training phase. Crucially, the encoder weights were frozen, and no supervised classifier was trained for these new classes.

**Table 4** Error rate (%) of DAS-MAEv2 in practical external damage prevention application. The best results are in bold.

| Training data | DAS-MAE | | Scratch model | RI | |
|---|---|---|---|---|---|
| (portion\|volume) | v1 [23] | v2 | (DAS-MAEv2 structure) | (Scratch, DAS-MAEv2) | (DAS-MAEv1, DAS-MAEv2) |
| 100\|1100 | 5.0 | **4.7** | 19.1 | 75.4 | 6.0 |
| 62.5\|700 | 12.4 | **5.3** | 28.2 | 81.2 | 57.3 |
| 25\|280 | 16.2 | **9.1** | 34.8 | 73.9 | 43.8 |

The t-SNE projects the learned representations into a 2D space for visualization. To ensure comparability, the t-SNE hyperparameters (e.g., perplexity, learning rate, and number of iterations) were kept identical to those used in Subsection 3.1. Figure 4(b) presents the t-SNE visualization of the 8 novel event types. Despite having zero prior exposure to these events, the model exhibits remarkable representation learning capability. As observed, samples belonging to different novel categories are well-separated in the latent space. Samples within the same category tend to cluster tightly, indicating that the model captures the consistent underlying signatures of each event type. Although minor overlaps exist between semantically similar events, the overall distinctive clustering confirms that DAS-MAEv2 does not merely memorize training labels. Instead, it learns generic and robust signal representations, making it highly effective for identifying new threats in open-set industrial scenarios.

### 3.4 Field evaluation of representation generalization

The pre-trained DAS-MAEv2 is fine-tuned and applied to the external-damage-prevention dataset. Notably, this data distribution exhibits significant and intentional class imbalance that reflects real-world operational conditions. Excavator-related activities (Classes 3 and 4) occur nearly 10 times more frequently than electric drill operations (Class 5), while background noise recordings (Class 0) are particularly sparse at just 42 samples. This carefully constructed imbalance serves two critical evaluation purposes: First, it rigorously tests the model's ability to maintain detection sensitivity for rare but operationally important events (such as electric drill operations). Second, it challenges the learned representations to resist classification bias toward dominant categories (like excavator activities) that could otherwise lead to overfitting. While this class imbalance presents substantial learning challenges, we deliberately retained the original data distribution to accurately reflect real-world operational conditions without employing any class balancing strategies. This approach provides a rigorous evaluation framework that authentically assesses the model's capacity to develop robust discriminative features. It is a critical capability for practical field deployment where rare but critical events must be reliably detected.

The waterfall plot dataset was partitioned into training and testing sets using a 4:1 ratio per event class. We evaluated the pre-trained DAS-MAE through fine-tuning on three subsets of the training data: full capacity (100%, ∼1100 samples), medium capacity (62.5%, ∼700 samples), and low capacity (25%, ∼280 samples), with each configuration trained for 150 epochs. The results are given in Table 4. When utilizing the complete training set, the pre-trained model achieved a 4.7% test ER, exceeding the empirically established 10% ER threshold for industrial fiber sensing deployments by 5.3 percentage points. This error rate translates to approximately 47 daily classification errors for a system monitoring 1000 daily events. This superior performance represents a 75.4% relative improvement (14.4% absolute improvement) compared to scratch-model training (trained until convergence) and a 6% relative improvement over DAS-MAEv1. As evidenced in Figure 5(a), the pre-trained DAS-MAEv2 exhibits superior performance across all classes. It commits significantly more corrections (7 vs. 1) in full-load forklift operation classification (Class 6), demonstrating effective mitigation of class imbalance. Our model reduces false alarms to essentially alleviate alarm fatigue for operators and prevent unnecessary field deployments (saving operational costs). The pre-trained DAS-MAEv2 resolves the confusion between excavator states (excavation vs. motion) that plagues the scratch model. Minimizing missed detections is paramount as it ensures a high probability of detecting true threats, directly enhancing infrastructure integrity and safety. Under data-limited conditions, the pre-trained model maintained robust performance with error rates of 5.3% (62.5% training data) and 9.1% (25% training data), consistently outperforming scratch models by >70% relative improvement (∼20% absolute improvement) and DAS-MAEv1 by >40% relative improvement. Figures 5(b) and (c) illustrate the specific confusion matrices. These results demonstrate the substantial effectiveness of DAS-MAEv2's pre-training phase and prove the high transferability of our model in different applications with novel event types.

### 3.5 Model robustness to noise

To systematically evaluate the robustness of DAS-MAEv2 against environmental factors, we conducted a sensitivity analysis using our external-damage-prevention dataset. Gaussian white noise of varying intensities is introduced to the dataset to simulate different physical conditions, resulting in signal-to-noise ratios (SNR) of 0.001, 0.01,

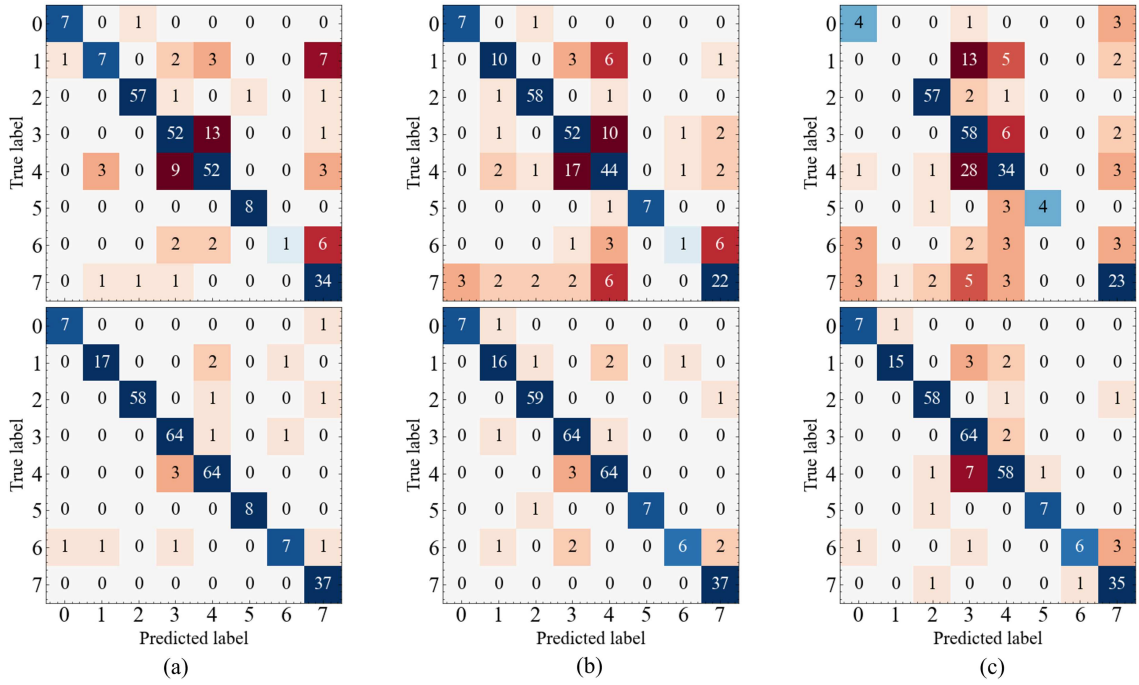


**Figure 5** (Color online) Confusion matrix comparison between the pre-trained model (i.e., DAS-MAEv2) and the scratch model across different training set sizes. The upper row in each comparison displays results from the scratch model (same architecture but trained from scratch), while the bottom row shows corresponding results from DAS-MAEv2. (a) 1100 training samples: DAS-MAEv2 reduces the error rate (ER) from 19.1% (scratch model) to 4.7%. (b) 700 training samples: DAS-MAEv2 reduces ER from 28.2% to 5.3%. (c) 280 training samples: DAS-MAEv2 reduces ER from 34.8% to 9.1%.

0.1, 1, and 10. Notably, the original dataset without added noise is defined as having an infinite SNR (SNR = $\infty$). We utilized the pre-trained DAS-MAEv2 to extract representations from data at these different SNR levels and visualized the distributions via t-SNE (same hyperparameters as in Subsection 3.1), as shown in Figure 6. Simultaneously, to quantify the impact of noise on model performance, we fine-tuned the pre-trained model for 150 epochs on the training sets corresponding to each SNR level. The resulting 8-class classification error rates are reported in the corresponding t-SNE plots. As observed in Figure 6(b), at low noise levels (SNR = 10), the dimensionality reduction results closely resemble those of the noise-free data in Figure 6(a). Events of the same category remain tightly clustered while distinct types are well-separated, with both scenarios achieving a low error rate of 4.7%. However, as noise increases to SNR = 1 and 0.1, the learned representation is partially affected. In Figures 6(c) and (d), the clustering density of intra-class events decreases, and the fine-tuned error rates rise to 5.3% and 6.9%, respectively.

When the noise intensity is further increased to SNR = 0.01 in Figure 6(e), the representations begin to degrade significantly. While the road roller in compaction class (orange) remains distinguishable with a clear boundary, the representations of the other seven event types become mixed. Although fine-tuning improves the error rate to 15.6%, this performance level may fall short of strict practical application requirements. Finally, at the extreme noise level of SNR = 0.001 in Figure 6(f), the features become completely entangled, making it difficult to identify distinct class boundaries. Even after fine-tuning, the model yields an error rate of 52.6%, indicating a failure to classify correctly under such extreme noisy conditions. Overall, these results demonstrate that DAS-MAEv2 possesses strong robustness against noise, maintaining high performance and clear feature separation even under significant noise interference (down to SNR = 0.1).

## 4 Ablation study

To optimize DAS-MAEv2's configuration for waterfall plot analysis and representation learning, we performed hyperparameter ablation studies. Using a controlled experimental protocol, we first pre-trained models with individual hyperparameter variations under identical training schedules (Section 3). The learned representation quality is then assessed through standardized linear probing and fine-tuning the pre-trained model on the benchmark dataset [26]. The test-set classification error rate serves as the quantitative metric. Throughout these experiments, we strictly

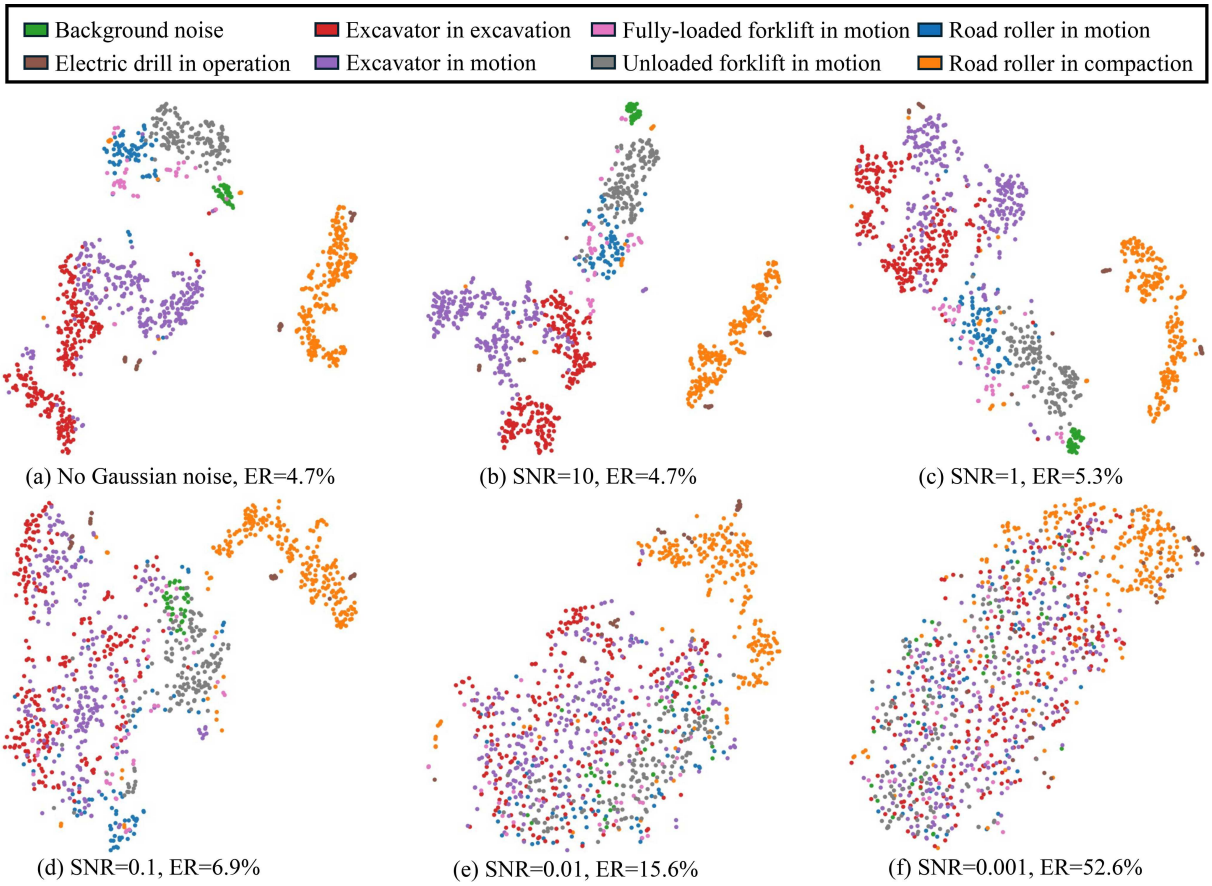


**Figure 6** (Color online) t-SNE visualization of DAS-MAEv2 representations and robustness analysis under different signal-to-noise ratio (SNR) conditions. ER reported in each subplot represents the 8-class classification error rate obtained after fine-tuning the pre-trained model on the training set with the corresponding SNR level.

modified only one hyperparameter at a time while keeping all other architectural components fixed to ensure isolated variable analysis.

### 4.1 STFT transformation and video pre-training

To thoroughly investigate the impact of STFT transformation and video pre-training on model performance, we conducted a detailed comprehensive comparison across distinct model configurations: (1) the original DAS-MAEv1 [23], (2) a modified version of DAS-MAEv1 using DAS-MAEv2's hyperparameter settings on Transformer blocks for the same model size as DAS-MAEv2 (denoted as v1-same-params), (3) DAS-MAEv2 architecture without video pre-training (v2-w/o-video), and (4) the complete DAS-MAEv2 with all enhancements. The t-SNE visualizations of learned representations from these models on the open dataset are presented in Figure 7. The visualization results reveal important insights about the models' representation learning capabilities. In Figure 7(a), the original DAS-MAEv1 shows well-clustered representations where samples from the same event class aggregate together with clear separation between different classes. By contrast, the v1-same-params variant in Figure 7(b) demonstrates degraded clustering performance, with noticeable overlaps between distinct event classes, particularly between digging (orange) and walking (brown). This performance degradation was expected and can be explained by the fundamental architectural differences between DAS-MAEv1 and DAS-MAEv2. The DAS-MAEv1 architecture was specifically optimized for capturing spatial-temporal coherence in waterfall plots, while DAS-MAEv2's parameter configuration was designed to better handle joint spatial-temporal-frequency coherence. The benefits of STFT transformation become evident in Figure 7(c), where the v2-w/o-video model shows obvious improvement over both DAS-MAEv1 variants. The representations exhibit clearer class boundaries than DAS-MAEv1, with distinct separation emerging between digging (orange) and walking (brown) events, as well as between shaking (purple) and watering (red) events. This demonstrates the advantage of incorporating frequency-domain information through STFT transformation. When examining the complete DAS-MAEv2 in Figure 7(d), we observe that the additional video pre-training provides extra benefits. The most noticeable improvement appears in the background noise class,

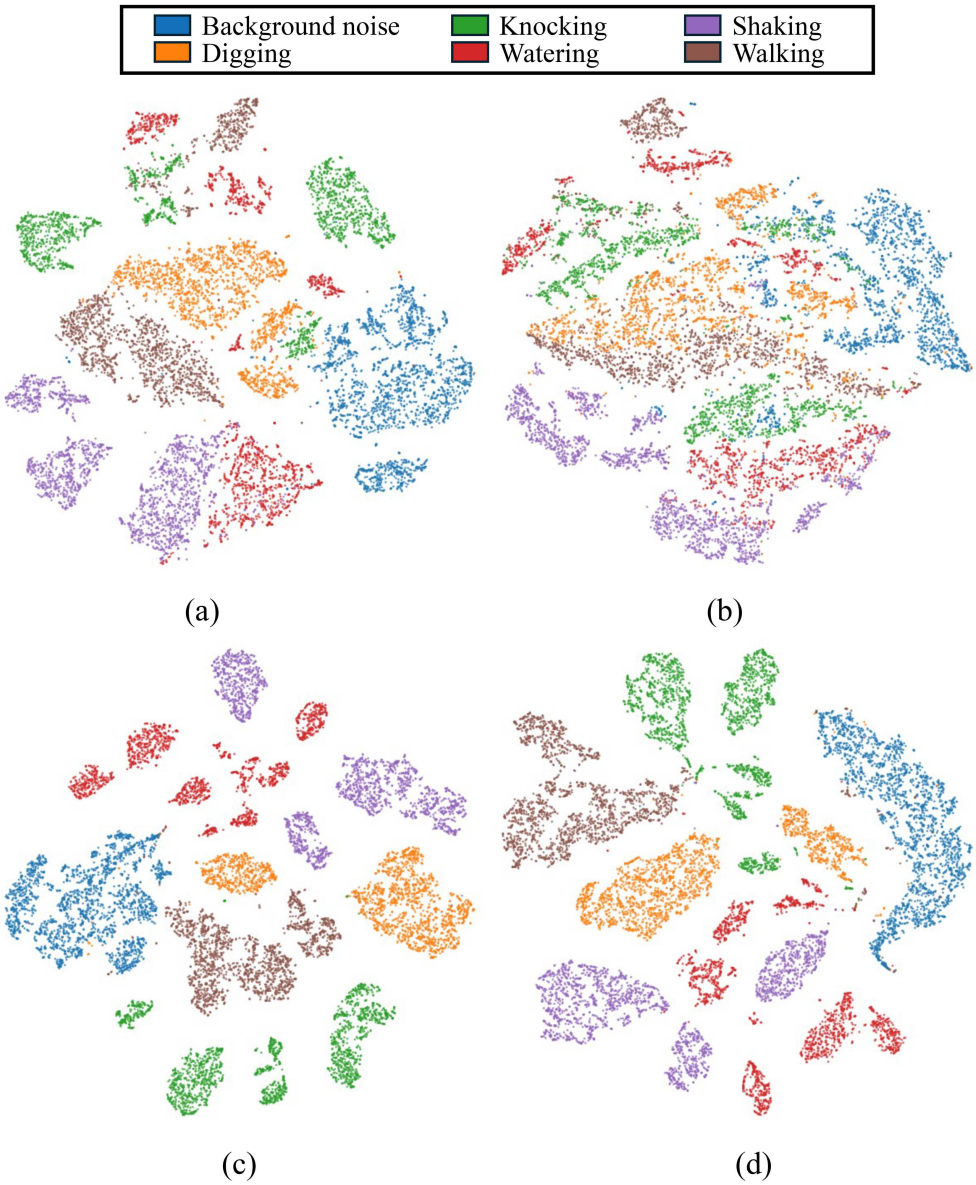


**Figure 7** (Color online) The t-SNE visualization of representations on the open dataset [26]. Labels are used for color assignment to evaluate the quality of learned representations. (a) Representations from DAS-MAEv1; (b) representations learned by v1-same-params (a modified version of DAS-MAEv1 using DAS-MAEv2's hyperparameter settings); (c) representations obtained from v2-w/o-video (DAS-MAEv2 architecture without video pre-training); (d) representations from DAS-MAEv2.

**Table 5** Ablation study (%) on STFT transformation and video pre-training. The best results are in bold.

| DAS-MAE | Mask ratio | ER of linear probing | ER of fine-tuning |
|---|---|---|---|
| DAS-MAEv1 from [23] | 50 | 2.45 | 0.32 |
| v1-same-params | 50 | 7.86 | 2.77 |
| v2-w/o-video | 90 | 0.42 | 0.16 |
| DAS-MAEv2 | 90 | **0.23** | **0.06** |

whose representations become more tightly clustered. For other event classes, the enhancements from video pre-training are less observable. Collectively, these visualization results demonstrate that STFT transformation and video pre-training contribute to learning more discriminative representations.

Quantitative results are presented in Table 5. Our experimental results show that the original DAS-MAEv1 architecture achieved error rates of 2.45% for linear probing and 0.32% for fine-tuning under the optimal mask ratio of 50%. When using the v1-same-params model, we observed increased ER of 7.86% for linear probing and 2.77%

**Table 6** Ablation study (%) on mask ratio. The best results are in bold.

| Mask ratio | ER of linear probing | ER of fine-tuning |
|---|---|---|
| 70% | 0.35 | 0.13 |
| 80% | 0.48 | **0.06** |
| 90% | **0.23** | **0.06** |
| 95% | 1.10 | 0.19 |
| 98% | 11.33 | 0.51 |

for fine-tuning. This result aligns with the observation in the t-SNE visualization in Figures 7(a) and (b). The implementation of STFT transformation in the v2-w/o-video model brought substantial performance improvements, reducing the error rates to 0.42% for linear probing and 0.16% for fine-tuning. These results represent significant relative improvements of 94.7% and 94.2%, respectively, compared to the v1-same-params configuration. During our experiments, we discovered that the optimal mask ratio for the DAS-MAEv2 architecture increased to 90%, which suggests that the spatial-temporal-frequency 'waterfall plots' exhibit much higher information redundancy compared to raw spatial-temporal waterfall plots. We believe this enhanced redundancy stems from STFT's unique ability to explicitly reveal the critical frequency-domain features hidden in the raw data. When we further incorporated video pre-training to create the full DAS-MAEv2 model, we achieved our best results with error rates of 0.23% for linear probing and 0.06% for fine-tuning. These figures correspond to additional relative improvements of 45.2% and 62.5% over the v2-w/o-video configuration, clearly demonstrating the complementary benefits of both architectural innovations. The STFT transformation provides superior frequency-aware representation learning capabilities, while the video pre-training contributes valuable transferable feature extraction abilities that further enhance model performance.

### 4.2 Mask ratio

The mask ratio constitutes a critical hyperparameter in DAS-MAEv2 that regulates the difficulty of the reconstruction task and determines the quality of learned representations. This parameter requires careful balancing, where the task must be sufficiently challenging to prevent the model from relying on simple interpolation strategies, yet not so difficult as to hinder the learning of fundamental signal characteristics. This delicate balancing ensures the encoder acquires precisely the necessary information to learn high-quality representations that capture meaningful semantics (e.g., event categories). As demonstrated in Table 6, the optimal representations emerge at an aggressive 90% mask ratio, yielding minimal error rates of 0.23% (linear probing) and 0.06% (fine-tuning). Both more conservative (less than 70%) and excessively aggressive (over 95%) masking ratios result in suboptimal representations. The model exhibits different sensitivity to mask ratios depending on the evaluation approach. Fine-tuning demonstrates robust performance across a wide range of mask ratios (70%–95%) in this classification task. In contrast, linear probing shows huge error rate reductions of 0.8% between 90% and 95% mask ratio, and 10% between 95% and 98% ratio. This contrast underscores the practical advantage of fine-tuning for achieving stable performance in practical applications.

The 90% optimal mask ratio for 3D spatial-temporal-frequency 'waterfall plots' presents a notable contrast to the 50% optimal ratio observed for 2D spatial-temporal waterfall plots. This difference indicates that our model can effectively leverage more heavily masked inputs while actually improving performance. Previous studies prove that the optimal mask ratio correlates strongly with the inherent redundancy of the input signal. For instance, video data (with higher redundancy than images) shows optimal mask ratios of 90% [38] compared to 75% for images [25]. Similarly, while 2D waterfall plots achieve best performance at 50% masking, 3D waterfall plots peak at 90%, suggesting significantly higher information redundancy in the latter case. Crucially, this finding implies that STFT successfully exposes latent information within 2D waterfall plots, making the signal appear more redundant to the DAS-MAEv2 model. This phenomenon provides evidence for the effectiveness of incorporating STFT transformation in our framework, as it enables the model to work with less visible inputs while simultaneously improving representation quality.

### 4.3 Mask sampling strategy

Table 7 systematically evaluates the impact of four distinct mask sampling strategies on DAS-MAEv2's representation learning performance. Visual illustrations of each strategy are provided in Figure 8. The spatial sampling strategy (Figure 8(b)) removes entire spatial-wise tubes and compels the model to perform reconstruction using spatial-adjacent temporal-frequency information (i.e., time-series similarity in the waterfall plot). This design

**Table 7** Ablation study (%) on mask strategy. The best results are in bold.

| Mask strategy | ER of linear probing | ER of fine-tuning |
|---|---|---|
| Random sampling | **0.23** | **0.06** |
| Spatial sampling | 4.48 | 1.01 |
| Temporal sampling | 4.09 | 0.48 |
| Frequency sampling | 4.03 | 0.41 |

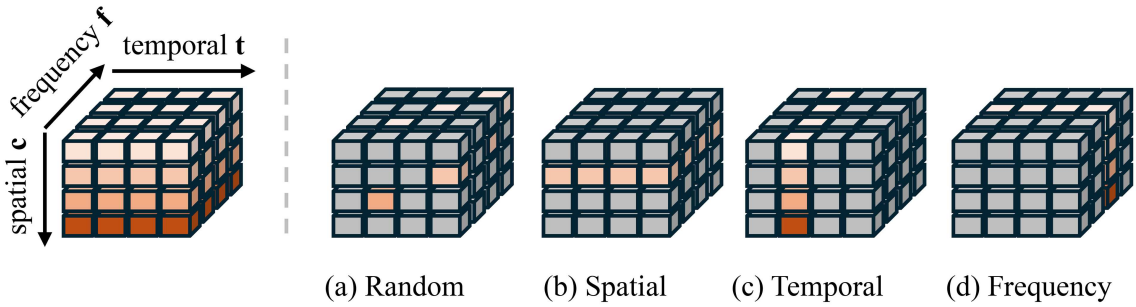


**Figure 8** (Color online) Mask sampling strategies determine the difficulty of the pre-training reconstruction task, influencing the quality of learned representations. (a) Random sampling (our default) masks the tubes following a uniform distribution; (b) spatial sampling removes the entire spatial-wise tubes; (c) temporal sampling removes the entire temporal-wise tubes; (d) frequency sampling removes the entire frequency-wise tubes.

facilitates learning inter-channel relationships across 2D temporal-frequency data, but it presents challenges for capturing intrinsic temporal-frequency distributions within individual channels. Under this limitation, DAS-MAEv2 only achieved error rates of 4.48% (linear probing) and 1.01% (fine-tuning) with spatial sampling, highlighting the importance of spatial correlations in waterfall plot analysis. These results underscore the advantages of distributed sensors over traditional single-point sensors. The spatial correlations enable cross-validation between adjacent sensors, provide complementary signal features unavailable in isolated measurements, and allow for error correction through majority consensus among neighboring channels.

The temporal sampling strategy (Figure 8(c)), which eliminates temporal redundancy by masking temporal-wise tubes, demonstrated improved performance with a 4.09% linear probing error rate (representing a 9% relative improvement over spatial sampling) and 0.48% fine-tuning error rate (52% relative improvement). Frequency sampling (Figure 8(d)), which removes entire frequency-wise tubes, achieved further gains with a 4.03% linear probing error rate (1% relative improvement over temporal sampling) and 0.41% fine-tuning error rate (15% relative improvement). These progressive improvements confirm the dominant role of frequency information in 3D waterfall plot representation learning, where accurate understanding of frequency distributions enables higher-quality feature extraction.

The random sampling strategy in Figure 8(a) proved to be the optimal approach, achieving superior performance with a 0.23% linear probing error rate (94% relative improvement over frequency sampling) and 0.06% fine-tuning error rate (85% relative improvement). The effectiveness of random sampling stems from its ability to simultaneously leverage spatial, temporal, and frequency correlations while introducing beneficial stochasticity that enhances model robustness and prevents overfitting. This comprehensive sampling approach allows the model to learn balanced representations that capture all relevant dimensions of the 3D waterfall plot data structure.

### 4.4 STFT output data format

The STFT transformation converts the original 2D waterfall plot data into 3D complex-valued data. To accommodate our model's requirement for real-valued input, we evaluate three different real-valued formats of STFT results: magnitude-only (absolute values), concatenated magnitude and phase components, and concatenated real and imaginary parts. The concatenation operation is performed along the additional input dimension $D_i$. We conducted a systematic ablation study to evaluate how these formats affect DAS-MAEv2's representation learning capability. The comparative results for different input representations are presented in Table 8. When using concatenated real and imaginary parts as input, the model achieved the worst performance with linear probing and fine-tuning error rates of 76.01% and 81.22%, respectively. The result indicates the model failed to learn meaningful representations from this input format. In contrast, using concatenated magnitude and phase components yielded significantly better results, with error rates of 0.81% (linear probing) and 0.16% (fine-tuning). This improvement aligns with established knowledge in time-series signal processing (e.g., speech recognition), where spectral magnitude typically

**Table 8** Ablation study (%) on input data format. The best results are in bold.

| Data | ER of linear probing | ER of fine-tuning |
|---|---|---|
| Magnitude | **0.23** | **0.06** |
| Magnitude & Phase | 0.81 | 0.16 |
| Real & Imaginary | 76.01 | 81.22 |

contains more discriminative information than phase components. Since waterfall plots essentially represent spatially distributed time-series signals, the performance difference between these input formats can be explained by using this conclusion. With magnitude-phase concatenation, the model can selectively focus more on the magnitude component while largely ignoring the less relevant phase information. On the contrary, with real-imaginary concatenation, the magnitude and phase information become entangled in a joint probability distribution. This joint probability distribution makes using the discriminative magnitude features more challenging for the model. Further improvement was achieved when using only magnitude information as input. It yields the best performance with a linear probing error rate of 0.23% (72% relative improvement over magnitude-phase concatenation) and a fine-tuning error rate of 0.06% (63% relative improvement). These results demonstrate that the magnitude spectrum alone provides the most critical and discriminative information for event classification in DAS applications.

## 5 Discussion

Despite the superior performance demonstrated in our experiments, we acknowledge several limitations when transitioning this method to large-scale industrial deployments. Firstly, the Transformer's quadratic complexity poses challenges for edge-device inference, requiring future exploration of model compression. Secondly, while self-supervised learning reduces annotation costs, scaling to terabyte-scale datasets demands substantial computational resources. Thirdly, environmental robustness against complex non-Gaussian noise and domain shifts requires further validation under rigorous generalization protocols to ensure reliability across diverse industrial settings. These limitations highlight key directions for future research on efficient, scalable, and robust industrial adaptation.

## 6 Conclusion

In conclusion, this work presents DAS-MAEv2, an enhanced framework that significantly advances representation learning for DAS through two key innovations: STFT transformation for explicit frequency feature extraction and cross-domain video pre-training. By pioneering the first successful bridging of video sequences and waterfall plots in DAS applications, our approach establishes a new paradigm for learning highly transferable and capable representations that significantly outperform previous methods. The model's state-of-the-art performance validates video data as effective pre-training material, which enables learning comprehensive spatial-temporal-frequency representations. These capabilities prove particularly valuable for industrial applications requiring sophisticated waterfall plot analysis. Crucially, our framework enables data-efficient learning that aligns with IoT's requirements for flexible multimodal systems, reducing domain-specific data requirements through effective cross-modal transfer. This breakthrough not only advances DAS but also establishes a transformative approach to representation learning with broad implications for industrial IoT applications.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant Nos. 62435004, 62005163).

**References**

1 Rose K, Eldridge S, Chapin L, et al. The Internet of Things: an overview. Internet Society, 2015, 80: 1–53
2 Li S, Xu L D, Zhao S. The Internet of Things: a survey. Inf Syst Front, 2015, 17: 243–259
3 Masoudi A, Newson T P. Contributed review: distributed optical fibre dynamic strain sensing. Rev Sci Instruments, 2016, 87: 011501
4 He Z, Liu Q. Optical fiber distributed acoustic sensors: a review. J Lightwave Technol, 2021, 39: 3671–3686
5 Lu Y, Zhu T, Chen L, et al. Distributed vibration sensor based on coherent detection of phase-OTDR. J Lightwave Technol, 2010, 28: 3243–3249
6 Perol T, Gharbi M, Denolle M. Convolutional neural network for earthquake detection and location. Sci Adv, 2018, 4: e1700578
7 Ross Z E, Meier M, Hauksson E, et al. Generalized seismic phase detection with deep learning. Bull Seismological Soc Am, 2018, 108: 2894–2901
8 Tejedor J, Martins H F, Piote D, et al. Toward prevention of pipeline integrity threats using a smart fiber-optic surveillance system. J Lightwave Technol, 2016, 34: 4445–4453
9 Ho M, El-Borgi S, Patil D, et al. Inspection and monitoring systems subsea pipelines: a review paper. Struct Health Monitoring, 2020, 19: 606–645
10 Du C, Dutta S, Kurup P, et al. A review of railway infrastructure monitoring using fiber optic sensors. Sens Actuat A-Phys, 2020, 303: 111728
11 Rahman M A, Taheri H, Dababneh F, et al. A review of distributed acoustic sensing applications for railroad condition monitoring. Mech Syst Signal Processing, 2024, 208: 110983

12 Wu H, Wang C, Liu X, et al. Intelligent target recognition for distributed acoustic sensors by using both manual and deep features. Appl Opt, 2021, 60: 6878–6887
13 Yi J, Shang Y, Wang C, et al. An intelligent crash recognition method based on 1DResNet-SVM with distributed vibration sensors. Optics Commun, 2023, 536: 129263
14 Yang N, Zhao Y, Wang F, et al. Using phase-sensitive optical time domain reflectometers to develop an alignment-free end-to-end multitarget recognition model. Electronics, 2023, 12: 1617
15 Kayan C E, Yuksel Aldogan K, Gumus A. Intensity and phase stacked analysis of a Φ-OTDR system using deep transfer learning and recurrent neural networks. Appl Opt, 2023, 62: 1753–1764
16 Ma Y, Song Y, Song Q, et al. MI-SI based distributed optical fiber sensor for no-blind zone location and pattern recognition. J Lightwave Technol, 2022, 40: 3022–3030
17 Zhao X, Sun H, Lin B, et al. Markov transition fields and deep learning-based event-classification and vibration-frequency measurement for $\varphi$-OTDR. IEEE Sens J, 2021, 22: 3348–3357
18 Pan Y, Wen T, Ye W. Time attention analysis method for vibration pattern recognition of distributed optic fiber sensor. Optik, 2022, 251: 168127
19 van den Ende M, Lior I, Ampuero J P, et al. A self-supervised deep learning approach for blind denoising and waveform coherence enhancement in distributed acoustic sensing data. IEEE Trans Neural Netw Learn Syst, 2021, 34: 3371–3384
20 Li Z. Exploiting CNN-BiLSTM model for distributed acoustic sensing event recognition. In: Proceedings of International Conference on Artificial Intelligence and Communication (ICAIC 2024), 2024. 333–341
21 Wu H, Yang M, Yang S, et al. A novel DAS signal recognition method based on spatiotemporal information extraction with 1DCNNs-BiLSTM network. IEEE Access, 2020, 8: 119448
22 Li Z, Zhang J, Wang M, et al. Fiber distributed acoustic sensing using convolutional long short-term memory network: a field test on high-speed railway intrusion detection. Opt Express, 2020, 28: 2925–2938
23 Duan J, Chen J, He Z. DAS-MAE: a self-supervised framework for universal and high-performance representation learning of distributed acoustic sensing. J Lightwave Technol, 2026, 44: 2544–2556
24 Vaswani A. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017
25 He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 16000–16009
26 Cao X, Su Y, Jin Z, et al. An open dataset of $\varphi$-OTDR events with two classification models as baselines. Results Opt, 2023, 10: 100372
27 Wu H, Zhou B, Zhu K, et al. Pattern recognition in distributed fiber-optic acoustic sensor using an intensity and phase stacked convolutional neural network with data augmentation. Opt Express, 2021, 29: 3269–3283
28 Spica Z J, Ajo-Franklin J, Beroza G C, et al. PubDAS: a PUBlic distributed acoustic sensing datasets repository for geosciences. Seismological Res Lett, 2023, 94: 983–998
29 Wang C, Chen S, Wu Y, et al. Neural codec language models are zero-shot text to speech synthesizers. 2023. ArXiv:2301.02111
30 Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision. In: Proceedings of International Conference on Machine Learning, 2023. 28492–28518
31 Arnab A, Dehghani M, Heigold G, et al. Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 6836–6846
32 Ren M, Lu P, Chen L, et al. Theoretical and experimental analysis of O-OTDR based on polarization diversity detection. IEEE Photon Technol Lett, 2015, 28: 697–700
33 Chen D, Liu Q, He Z. 108-km distributed acoustic sensor with 220-p$\varepsilon/\sqrt{\text{Hz}}$ strain resolution and 5-m spatial resolution. J Lightwave Technol, 2019, 37: 4462–4468
34 Dong Y, Chen X, Liu E, et al. Quantitative measurement of dynamic nanostrain based on a phase-sensitive optical time domain reflectometer. Appl Opt, 2016, 55: 7810–7815
35 Fernández-Ruiz M R, Martins H F, Williams E F, et al. Seismic monitoring with distributed acoustic sensing from the near-surface to the deep oceans. J Lightwave Technol, 2022, 40: 1453–1463
36 Wu H, Chen J, Liu X, et al. One-dimensional CNN-based intelligent recognition of vibrations in pipeline monitoring with DAS. J Lightwave Technol, 2019, 37: 4359–4366
37 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. 2020. ArXiv:2010.11929
38 Tong Z, Song Y, Wang J, et al. Videomae: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 10078–10093
39 Loshchilov I. Decoupled weight decay regularization. 2017. ArXiv:1711.05101
40 Loshchilov I, Hutter F. Sgdr: stochastic gradient descent with warm restarts. 2016. ArXiv:1608.03983
41 Krizhevsky A, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2012. 25
42 Van der Maaten L, Hinton G. Visualizing data using t-sne. J Mach Learn Res, 2008, 9: 2579–2605
43 Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). Comput & Geosci, 1993, 19: 303–342
44 Li Y J, Cao X M, Ni W, et al. A deep learning model enabled multi-event recognition for distributed optical fiber sensing. Sci China Inf Sci, 2024, 67: 132404
45 Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 30
46 Li Y, Qin Y, Hu L, et al. An open-world semi-supervised recognition method (ODAS) for Φ-OTDR disturbance signals. IEEE Internet Things J, 2025, 12: 38307–38321
47 Kamanga I, Tang R, Wang Z, et al. Performance enhancement of Φ-OTDR event classification via dynamic MFCCs and multi-scale discriminator-based FastGAN data augmentation. Measurement, 2026, 257: 118927
48 Luo Z, Wu H, Ge Z, et al. Real-time event recognition of long-distance distributed vibration sensing with knowledge distillation and hardware acceleration. IEEE Internet Things J, 2025, 12: 17896–17905
49 Li Y J, Hu L Q, Yu K L. Unleashing potentials with deep learning: decoding the complex events for distributed fiber optic sensing applications. Sci China Inf Sci, 2024, 67: 159402
50 Jiang W, Jiang Y. St-t: a spatio-temporal transformer for phai-otdr multi-location time series classification. In: Proceedings of the 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2024. 1497–1502
51 Chen X, Yang C, Yu H, et al. Research on pattern recognition method for $\varphi$-OTDR system based on dendrite net. Electronics, 2023, 12: 3757