Pattern Recognition 183 (2027) 114997



Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr





Physics-constrained spatiotemporal modeling for temporal localization in structured multichannel acoustic signals

Long Wang [a,b], Xinming Wu [a,b,*], Zhifeng Sun [c], Yimin Dou [a,b]

[a] State Key Laboratory of Precision Geodesy, School of Earth and Space Sciences, University of Science and Technology of China, Hefei, 230026, Anhui, China
[b] National Field Observation and Research Station (Anhui Mengcheng) for Geophysics, University of Science and Technology of China, Mengcheng, 233500, Anhui, China
[c] China Oilfield Services Limited, Beijing, 101149, China

ARTICLE INFO

Keywords:
First arrival picking
Physics constrained
Synthetic datasets
Spatiotemporal modeling
Real time

ABSTRACT

Temporal localization in structured multichannel acoustic signals, such as first-arrival picking in geophysical logging, is a fundamental yet challenging task. Traditional signal processing methods rely heavily on manual heuristics, while existing data-driven deep learning approaches often overlook the inherent physical priors of multichannel arrays, resulting in a trade-off among accuracy, stability, and real-time performance in field applications. To address these challenges, we propose FA-Picker, a physics-constrained spatiotemporal modeling framework for first-arrival picking. This model explicitly incorporates the depth dimension to construct a three-dimensional input data volume, enabling the capture of joint spatiotemporal features across multiple receivers and consecutive measurements. Moreover, an end-to-end multi-task architecture is designed, where coarse first-arrival segmentation guides continuous arrival-time regression. A dual-constraint mechanism is introduced, combining soft loss-based guidance with hard physical constraints, ensuring that the localized temporal points strictly comply with the underlying wave propagation laws. Furthermore, to alleviate the scarcity of high-quality labeled data, we exploit the highly structured characteristics of multichannel acoustic arrays to develop a comprehensive synthetic data generation workflow that simulates diverse formation variations and acquisition environments. Experimental results on field datasets demonstrate the effectiveness of the proposed modeling and data augmentation strategies, showing the superior performance of FA-Picker in practical first-arrival picking tasks. The code is available at: https://github.com/longw0110/Synthetic-workflow-for-FA-Picker.

1. Introduction

As global oil and gas exploration and development continue to advance toward deeper formations, complex geological structures, and subtle lithologic traps, the fine-scale characterization of small-scale lithologic reservoirs has become a central focus of modern exploration activities. In this context, high-precision formation velocity models serve as a fundamental basis for seismic inversion and imaging. Array acoustic logging, acting as a critical bridge between borehole geophysics and surface seismic exploration, plays an indispensable role in this process [1]. The compressional and shear wave velocities derived from array acoustic logging are widely used in synthetic seismogram generation, lithology identification, fracture detection, and elastic parameter estimation [2,3]. By recording full-waveform data using multiple receivers, array acoustic logging tools enable the estimation of formation slowness (the reciprocal of velocity) through accurate picking of first-arrival times, as shown in Fig. 1. However, with the construction and rapid development of intelligent oilfields, the explosive growth of array waveform data has posed significant challenges to achieving real-time, accurate, and physically consistent first-arrival picking, making it a pressing problem in industrial applications [4].

Historically, first-arrival picking has primarily relied on traditional signal-processing-based approaches. Classical threshold-based methods are simple to implement but highly sensitive to noise. To mitigate this limitation, Willis and Toksöz [5] proposed a hybrid strategy that first obtains a coarse arrival using a threshold method and then refines the estimate through cross-correlation, aiming to improve robustness against noise. Exploiting the coherence characteristics of array signals, Kimball and Marzetta [6] introduced the slowness–time coherence (STC) method, which has been widely adopted in high-quality data processing and established as an industry standard. Nevertheless, when applied to low signal-to-noise ratio (SNR) data, these traditional approaches often require extensive manual parameter tuning [7]

* Corresponding author.
E-mail addresses: long_wang@mail.ustc.edu.cn (L. Wang), xinmwu@ustc.edu.cn (X. Wu), sunzf@cosl.com.cn (Z. Sun), douyimin@ustc.edu.cn (Y. Dou).

https://doi.org/10.1016/j.patcog.2026.114997
Received 26 March 2026; Received in revised form 19 August 2026; Accepted 19 September 2026
Available online 24 September 2026
0031-3203/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

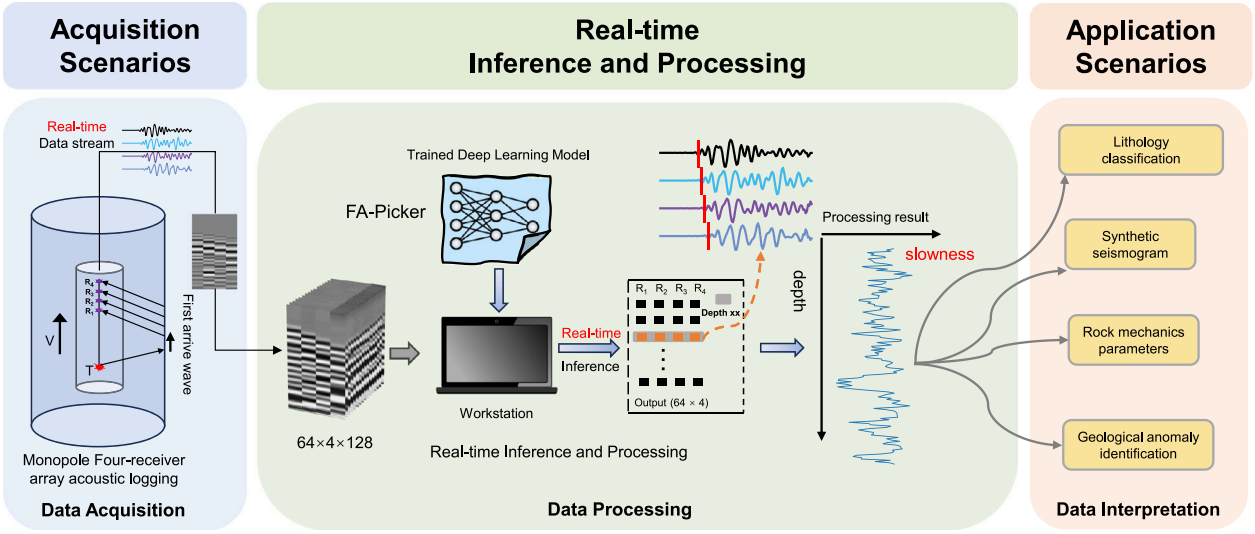


Fig. 1. Schematic of the intelligent real-time first-arrival picking workflow for array acoustic logging. (Left) Data acquisition, where a monopole four-receiver tool captures waveform data transmitted in real-time; (Middle) Data processing, where the FA-Picker processes batched samples (64 depth points) to derive formation slowness curves; and (Right) Engineering applications.

and are easily misled by structured noise events [8], such as interference waves arriving earlier than the true P-wave, resulting in spurious peaks and false picks. Evidently, expert-dependent manual or semi-automatic workflows can no longer meet the stringent real-time processing requirements imposed by large-scale modern logging operations.

In recent years, deep learning (DL) has emerged as a promising alternative for automated first-arrival picking due to its powerful feature extraction capability [9,10]. Some authors [11,12] applied DL models to P-wave picking in laboratory-scale acoustic emission experiments, demonstrating encouraging performance under controlled conditions. However, such laboratory datasets often fail to capture the complexity of real logging environments. For field acoustic logging data, some other researchers proposed DA-PhaseNet [13], which incorporates channel-wise and spatial attention mechanisms to enhance inter-trace time-delay features. In the broader seismic and microseismic fields, recent studies have further explored GAN-based data augmentation for low-SNR first-arrival picking [14], semi-supervised learning to reduce the dependence on manual labels [15], and graph neural networks for multi-station phase picking, association, and event location [16]. Related graph-based and physics-informed frameworks have also shown the potential of incorporating structural relationships, physical priors, or causal dependencies into data-driven models [17,18].

Despite these advances, existing methods still have limitations when applied to array acoustic logging first-arrival picking. Different from conventional seismic or microseismic phase picking, array acoustic logging involves a compact multi-receiver array, a short recording window, strong borehole noise, and continuous acquisition along the depth direction. The receiver arrivals should satisfy a physically consistent moveout order, while adjacent depth samples should vary smoothly because formation properties change continuously over short intervals [19]. However, most existing DL-based methods model single-depth waveforms as two-dimensional images or independent one-dimensional sequences [20,21], without explicitly exploiting depth-wise continuity or enforcing receiver-wise physical constraints. Therefore, robust and real-time first-arrival picking for array acoustic logging remains insufficiently studied. To address this gap, FA-Picker constructs a 3D depth–receiver–time input representation and combines inter-depth continuity modeling with a hard physical constraint module, thereby improving both prediction stability and geophysical consistency.

The generalization capability of deep learning models is known to depend heavily on the scale and diversity of training datasets, a phenomenon extensively validated in computer vision [22] and natural language processing [23]. However, in vertical geophysical domains such as array acoustic logging, acquiring large-scale, high-quality expert-labeled datasets is extremely challenging and costly [24]. Through detailed data analysis, we observe that array acoustic logging data exhibit strong inherent structural regularities [1,6], arising from the deterministic acquisition geometry and wave propagation physics. This intrinsic structure provides favorable conditions for generating physically consistent synthetic data [25] via forward modeling, thereby offering a viable solution to alleviate the scarcity of labeled training data and to support the development of generalizable deep learning models for industrial-scale applications.

To address these challenges, we propose FA-Picker, a physics-constrained, end-to-end deep learning framework for real-time first-arrival picking from large-scale array acoustic logging data. The main contributions of this work are summarized as follows:

1. We design a depth-aware three-dimensional input representation that introduces geological deposition priors by exploiting the continuous variation of acoustic waveforms along the depth direction.
2. We develop a physics-guided forward modeling and data augmentation pipeline to construct a high-quality labeled synthetic dataset representing diverse complex formation and acquisition conditions.
3. We propose an end-to-end multi-task learning framework, where segmentation implicitly models event detection and the regression module fuses segmentation features to capture long-range spatiotemporal dependencies and directly predict first-arrival times.
4. We conduct comprehensive performance evaluations, including blind-well tests, inference speed assessment, and comparative experiments, to demonstrate the superiority of the proposed method.

2. Related work

2.1. First-arrival picking

Array acoustic logging typically utilizes a multi-transmitter and multi-receiver configuration. As shown in Fig. 1, a monopole transmitter emits acoustic waves that undergo reflection and refraction at the

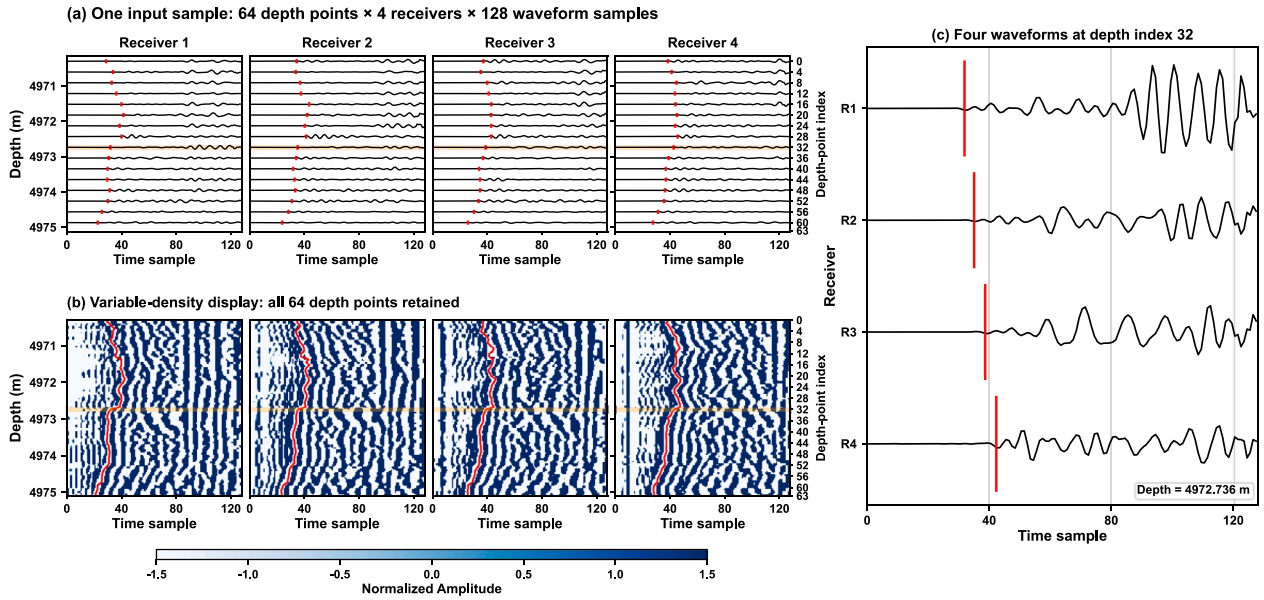


Fig. 2. Visualization of one input sample and its corresponding variable-density representation. (a) Decimated raw waveform display of a 64-depth-point input sample with four receivers and 128 time samples; only every fourth depth trace is shown for clarity. (b) Variable-density display of the same input sample, where all 64 depth points are retained. (c) Enlarged four-receiver waveform example at depth index 32. Red markers/curves denote first-arrival picks, and the orange band indicates the selected depth index.

borehole wall. At the critical incident angle, a refracted wave, termed a head wave, propagates along the interface between the formation and the borehole fluid. During its propagation along this interface, the head wave acts as a continuous secondary source, generating parallel reflections that subsequently reach each receiver [1]. Since the formation velocity is generally higher than that of the borehole mud, these waves reach the receivers earlier than other wave components and are therefore regarded as the first arrivals. First-arrival picking aims to determine the temporal position of the first arrival on each recorded waveform trace. As illustrated in Fig. 2(c), for a given depth point, the four receiver waveforms correspond to four first-arrival positions along the time-sample axis, marked by red indicators. Therefore, first-arrival picking can be viewed as a structured temporal localization task for multichannel acoustic waveforms.

2.2. Depth dimension addition

Formation properties usually vary smoothly over short depth intervals [19]. In array acoustic logging, the depth sampling interval is typically 0.1524 m or 0.0762 m, and the temporal sampling interval is 8 μs. Within such short depth intervals, the formation response and the corresponding first-arrival times are expected to exhibit strong depth-wise continuity. Therefore, instead of treating each depth point independently, FA-Picker explicitly incorporates adjacent depth information into the input representation.

As illustrated in Fig. 2(a), one input sample is constructed from 64 consecutive depth points. At each depth point, four receiver waveforms are recorded, and each waveform contains 128 temporal samples. Therefore, the input sample forms a three-dimensional data volume with a size of $64 \times 4 \times 128$, corresponding to Depth × Receiver × Time. In this representation, the depth dimension preserves the continuity of adjacent logging measurements, the receiver dimension describes the inter-receiver moveout pattern, and the time dimension represents the temporal waveform samples from which the first-arrival position is identified.

This 3D representation enables the network to learn two types of structured information simultaneously: the monotonic variation of first-arrival times across receivers at the same depth, and the smooth variation of first-arrival trajectories across adjacent depths. As shown in Fig. 2(b), the variable-density display further illustrates the depth-wise continuity of the first-arrival trajectories. Therefore, the 3D U-Net [26] can exploit inter-depth waveform continuity.

2.3. Segmentation and regression

Deep learning methodologies for arrival-time picking are essentially designed to learn nonlinear mappings from complex waveform inputs to temporal outputs [27,28]. Currently, most dominant paradigms treat arrival picking as an event detection task formulated via semantic segmentation [13,21]. By identifying sparse local regions associated with the first-arrival energy, these methods discretize continuous arrival times into binary or probabilistic masks, enabling the network to distinguish signal-dominant zones from background noise. Conversely, Transformer-based architectures have demonstrated superior capabilities in capturing long-range dependencies and cross-channel relationships [29,30], making them ideal for modeling correlated waveform sequences in array-based measurements.

Recent studies suggest that integrating complementary learning paradigms can significantly bolster model robustness, particularly under low SNR conditions [31]. Consequently, a hierarchical strategy is proposed: semantic segmentation provides structural priors by calibrating candidate arrival zones, while a regression module performs refined localization within these regions. This dual-path approach fuses the structured feature extraction of convolutional neural networks (CNNs) with the global modeling advantages of sequential models [32].

3. Synthetic datasets generation

Due to the array-based configuration of transmitter and receivers, the four raw waveforms observed at a single depth exhibit strong structural regularities. From these two perspectives — arrival time and inter-receiver time difference — the data demonstrate high physical manipulability.

Accordingly, diverse arrival scenarios can be efficiently synthesized via forward-modeling operations. Furthermore, to bridge the gap between synthetic data and field observations, we introduce a suite of physics-informed data augmentation techniques, including Signal Interval Suppression, Non-Stationary Baseline Drift and Stochastic Scattering Injection, which inject realistic noise, distortions and signal attenuation.

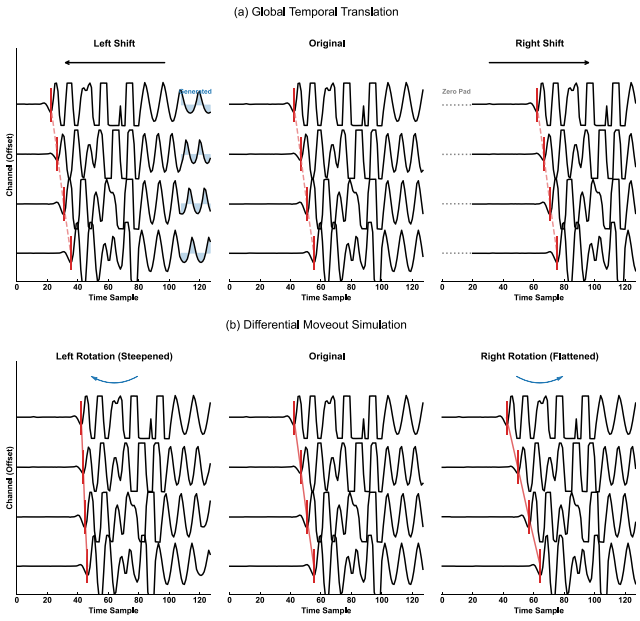


Fig. 3. Forward simulation of acoustic logging waveforms. (a) Global temporal translation. (b) Differential moveout simulation.

3.1. Forward simulation of acoustic logging waveforms

3.1.1. Global temporal translation

The specific position of the first arrival within the discrete time sequence ($t \in [0, 127]$) varies stochastically due to fluctuations in the instrument's acquisition trigger time and the propagation velocity of waves across different lithologies. To ensure the model maintains high detection accuracy across the entire temporal domain, Global Temporal Translation (GTT) is employed to simulate global delays or advances, guaranteeing full coverage of potential arrival positions, as shown in Fig. 3. We apply a uniform time shift Δt to the four-channel array waveform sequence $S(t)$. The augmentation logic is direction-dependent to maintain physical plausibility:

1. **Rightward Shift** ($\Delta t > 0$): To simulate a delayed arrival, we apply zero-padding to the start of the sequence to maintain a fixed length.
2. **Leftward Shift** ($\Delta t < 0$): To simulate an advanced arrival, the "future" data beyond the current window is unknown. To address this, we utilize a harmonic reconstruction method based on the Hilbert Transform. We extract the instantaneous phase $\phi(t)$ and dominant frequency to extrapolate the signal tail naturally, preventing non-physical discontinuities at the boundary [33].

The reconstructed signal tail $S_{\text{ext}}(t)$ for leftward shifts is modeled as:

$$S_{\text{ext}}(t) = \bar{A}\left[\cos(\hat{\phi}(t)) + \eta \sum_{j=1}^{J} j^{-\gamma} \cos((j+1)\hat{\phi}(t))\right] \tag{1}$$

where $\bar{A}$ is the mean envelope amplitude of the trailing window, η is a weighting factor (set to 0.3), J is the number of harmonics(set to 5), and γ is the decay rate(set to 1.5). The extrapolated phase $\hat{\phi}(t)$ is defined as:

$$\hat{\phi}(t) = \phi_{\text{end}} + \omega_{\text{dom}}(t - t_{\text{end}}) \tag{2}$$

This ensures that the augmented data maintains phase coherence with the original signal, preventing the network from learning artifacts at the boundaries.

3.1.2. Differential moveout simulation

Intuitively, the first arrivals of the four-channel array waveforms form a straight line across the channel index, as shown in Fig. 3. The slope of this line is directly related to the formation acoustic velocity: a steeper line indicates a smaller inter-receiver transit time (moveout) and thus higher formation velocity, while a flatter line indicates lower formation velocity. To simulate diverse formation velocity modes, we alter this slope by applying linearly increasing time shifts across the array channels.

To ensure high precision during this geometric transformation, we first perform 2× upsampling on the raw waveforms. We keep the first channel fixed and apply relative shifts to the subsequent channels. Let $S_k^{\text{up}}(t)$ denote the upsampled waveform of the kth receiver ($k = 1, 2, 3, 4$). The time shift δ_k applied to the kth channel is defined as:

$$\delta_k = (k-1) \cdot n \tag{3}$$

where n represents the number of shifted sample points (step). The augmented waveform $S'_k(t)$ is obtained by:

$$S'_k(t) = S_k^{\text{up}}(t \pm \delta_k) \tag{4}$$

This operation ensures that the first arrivals of four channels remain strictly collinear. The direction of the shift corresponds to specific changes in formation properties:

1. **Leftward Rotation (Steepened):** This reduces the inter-receiver transit time, visually "steepening" the slope (clockwise rotation). Since the time difference decreases, this simulates a higher velocity formation.
2. **Rightward Rotation (Flattened):** This increases the inter-receiver transit time, visually "flattening" the slope (counter-clockwise rotation). Since the time difference increases, this simulates a lower velocity formation.

By varying n, we can systematically generate synthetic data representing a continuous range of geological velocity models.

3.2. Data augmentation

3.2.1. Signal interval suppression

Let $x(t)$ denote the raw acoustic waveform and t_{arr} represent the ground-truth first arrival time. As shown in Fig. 4(a), we model formation absorption or coupling issues [34] by applying a time-varying gain function $\alpha(t)$, resulting in the attenuated signal $x_{\text{sis}}(t) = x(t) \cdot \alpha(t)$. The $\alpha(t)$ defines a strong suppression zone immediately preceding the arrival (controlled by factor γ_1) and a subsequent relaxation zone (controlled by factor γ_2):

$$\alpha(t) = \begin{cases} \gamma_1, & 0 \le t < t_{\text{arr}} + \Delta t_1 \\ \gamma_2, & t_{\text{arr}} + \Delta t_1 \le t < t_{\text{arr}} + \Delta t_2 \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where $\Delta t_1 = 10$ and $\Delta t_2 = 30$ sample points define the suppression zones.

3.2.2. Non-stationary baseline drift

To simulate low-frequency instrument drift, we superimpose a non-stationary trend $d(t)$ onto the suppressed signal $x_{\text{sis}}(t)$, as shown in Fig. 4(b). The drift is modeled as a sinusoidal function with randomized amplitude A_d and period T_d, yielding the drifted signal $x_{\text{nbd}}(t)$:

$$x_{\text{nbd}}(t) = x_{\text{sis}}(t) + A_d \sin\left(\frac{2\pi t}{T_d} + \phi\right) \tag{6}$$

where $\phi \in [0, 2\pi)$ is a random phase. To ensure the drift manifests as a background trend rather than high-frequency noise, the period T_d is constrained to be a multiple of the waveform length L.

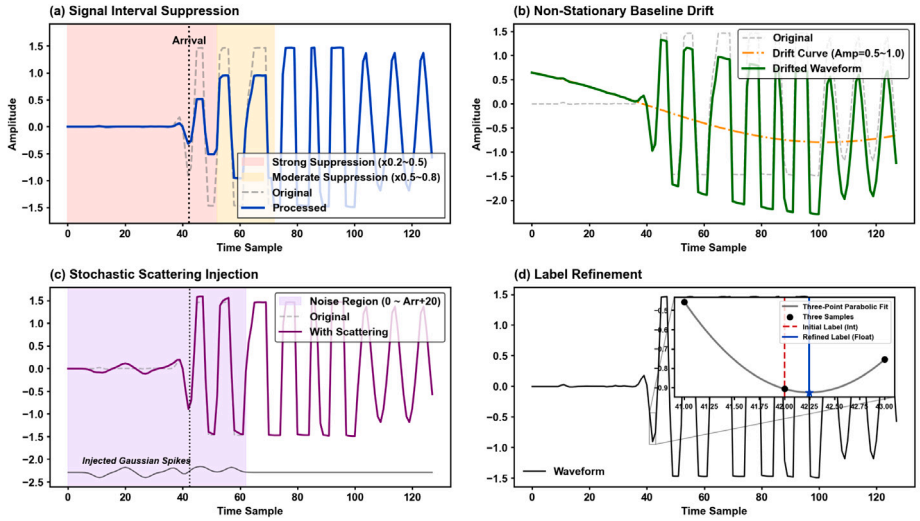


Fig. 4. Data augmentation. (a) Signal interval suppression. (b) Non-Stationary baseline drift. (c) Stochastic scattering injection. (d) Label refinement.

Table 1
Summary of parameter configurations for the four synthetic datasets.

Dataset	Signal interval suppression		Baseline drift		Scattering noise	
ID	Strong (γ_1)	Moderate (γ_2)	Amp. (A_d)	Period (T_d/L)	Level (β)	Count (K)
Syn_1	0.20–0.50	0.50–0.80	0.50–1.00	1.50–3.00	0.10	8
Syn_2	–	–	–	–	0.25	16
Syn_3	0.20–0.50	0.50–0.80	–	–	0.10	8
Syn_4	0.15–0.25	0.50–0.80	–	–	–	–

Note: "–" indicates the corresponding augmentation step is not applied.

3.2.3. Stochastic scattering injection

To mimic formation inhomogeneities [35], we inject stochastic scattering noise $n_{\text{scat}}(t)$ targeted at the first arrival region, as shown in Fig. 4(c). This component is modeled as a superposition of K localized Gaussian kernels scaled by the total signal energy E_x:

$$n_{\text{scat}}(t) = \beta\sqrt{E_x}\sum_{k=1}^{K} c_k \exp\left(-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right) \tag{7}$$

where β is the noise level coefficient, and $c_k \sim \mathcal{U}(0.03, 0.15)$ represents the scattering amplitude. The positions μ_k are randomly sampled within the targeted window, and the widths $\sigma_k \in [1, 3]$ simulate the diverse frequency content of scattered waves. The final augmented waveform is given by $x_{\text{aug}}(t) = x_{\text{nbd}}(t) + n_{\text{scat}}(t)$.

3.2.4. Label refinement

After noise injection, the original integer label t_{arr} may deviate slightly from the true trough position. To improve label precision, we apply three-point parabolic interpolation around the initial arrival label, as shown in Fig. 4(d). Using the amplitudes at t_{arr} and its two neighboring samples, the final target value $\hat{t}_{\text{arr}}$ is computed as:

$$\hat{t}_{\text{arr}} = t_{\text{arr}} + \frac{f[t_{\text{arr}}-1] - f[t_{\text{arr}}+1]}{2\left(f[t_{\text{arr}}-1] - 2f[t_{\text{arr}}] + f[t_{\text{arr}}+1]\right)}. \tag{8}$$

This interpolation estimates the parabola vertex, providing a sub-sample correction that better aligns the final target value with the local waveform trough.

Algorithm 1: Synthetic dataset generation and screening workflow

Input: Original labeled dataset D_0; maximum inter-receiver increment $\delta_{\max}$.
Output: Final synthetic training dataset D_{train}.

Generate synthetic samples using Global Temporal Translation and Differential Moveout Simulation;
Merge all generated samples into D_{syn};
for *each sample* $s \in D_{\text{syn}}$ **do**
  Calculate the three adjacent inter-receiver arrival-time differences;
  Discard s if any arrival-time difference is negative;
  Discard s if the difference between adjacent arrival-time differences exceeds 1 sample;
Apply the four augmentation configurations in Table 1 to obtain Syn_1–Syn_4;
Merge Syn_1–Syn_4 into D_{aug};
for *each sample* $s \in D_{\text{aug}}$ **do**
  Calculate the three adjacent inter-receiver arrival-time differences;
  Discard s if any arrival-time difference exceeds $\delta_{\max}$;
  Otherwise, add s to D_{train};
return D_{train};

3.3. Synthetic dataset configuration

The original real labeled dataset used for synthetic data generation consists of 74,147 depth samples from seven different wells, with high-quality first-arrival labels obtained through manual inspection and annotation. Based on this real labeled dataset, the synthetic dataset construction workflow is summarized in Algorithm 1. Large-scale array acoustic samples are first produced by GTT and Differential Moveout Simulation, followed by an initial inter-receiver consistency check. The remaining samples are augmented using the four configurations in

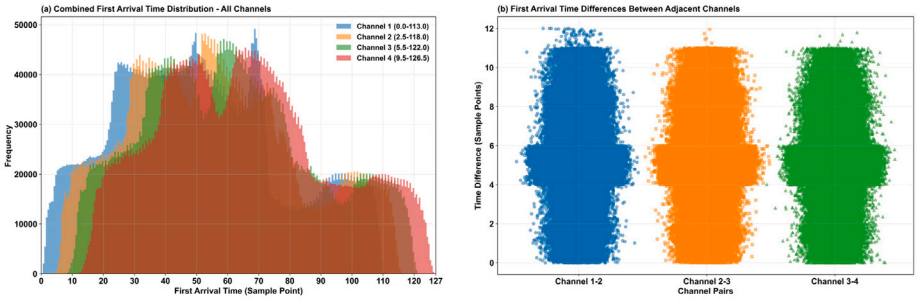


Fig. 5. Train datasets distribution. (a) Combined first arrival time distribution. (b) First arrival time difference between adjacent channels.

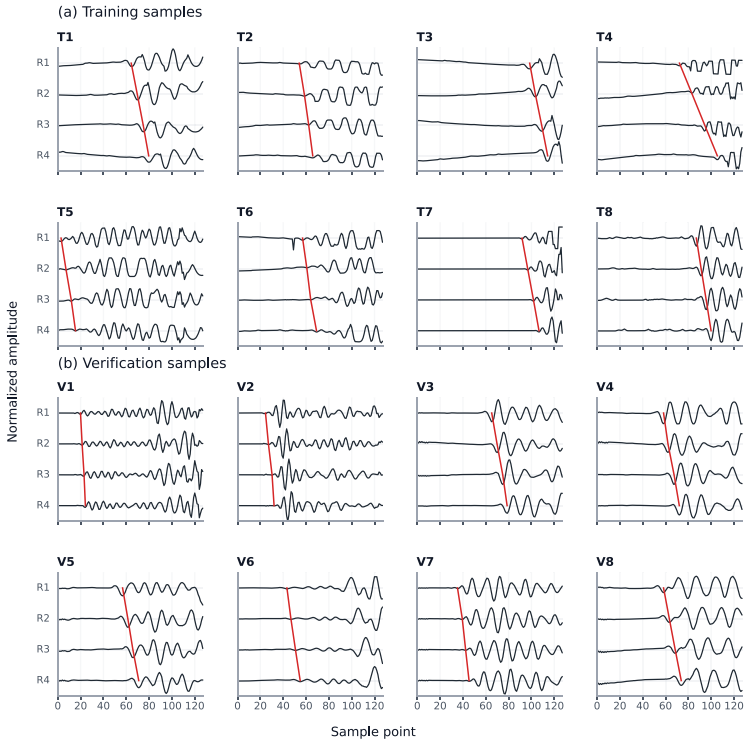


Fig. 6. Representative synthetic training and field verification samples. In each subplot, the red line connects the first-arrival positions of four receivers, showing the inter-receiver moveout pattern. The synthetic samples preserve field-like waveform characteristics while covering more challenging noise conditions.

Table 1 to obtain Syn_1–Syn_4, which cover different noise, baseline drift, and signal suppression conditions. A final $\delta_{\max}$-based filtering step is then applied to match the training-label distribution with the feasible prediction range of the PCM. Finally, 6.12 million labeled samples are retained for model training.

The characteristics of the training data are illustrated in Fig. 5 from the perspectives of arrival time and arrival time difference. As shown in Fig. 5(a), the first-arrival labels of the four receiver channels exhibit broad empirical distributions along the time axis. The samples are mainly concentrated in the middle portion of the recording window, while fewer labels appear near the temporal boundaries. Furthermore, Fig. 5(b) demonstrates that the distributions of the three arrival time differences are balanced, which is consistent with the physical property that arrival time differences remain identical or approximate across the array.

3.4. Visual comparison between synthetic and field samples

As shown in Fig. 6, the synthetic training samples preserve the main waveform characteristics of the field verification samples while introducing additional variations. Some synthetic samples exhibit stronger noise interference, weaker first-arrival events, waveform distortion, and larger inter-receiver arrival-time differences. In each subplot, the red line connects the first-arrival positions of four receivers, illustrating the inter-receiver moveout pattern. These observations suggest that the synthetic data are not restricted to idealized cases, but provide a diverse set of waveform conditions.

4. First arrival picker net

4.1. Dual branch and multitask-learning

Fig. 7 shows the architecture of FA-Picker, which adopts a unified multi-task learning framework [36] consisting of a first-arrival mask segmentation branch and a precise arrival-time regression branch. In the segmentation branch, first-arrival picking is formulated as a semantic segmentation task. Specifically, each waveform trace is treated as a sequence of temporal samples, where the coarse first-arrival position is labeled as the target class and the remaining samples are labeled as background. In this way, the segmentation label forms a binary mask, with the first-arrival region assigned to 1 and the background assigned to 0. The segmentation branch provides a coarse localization of the

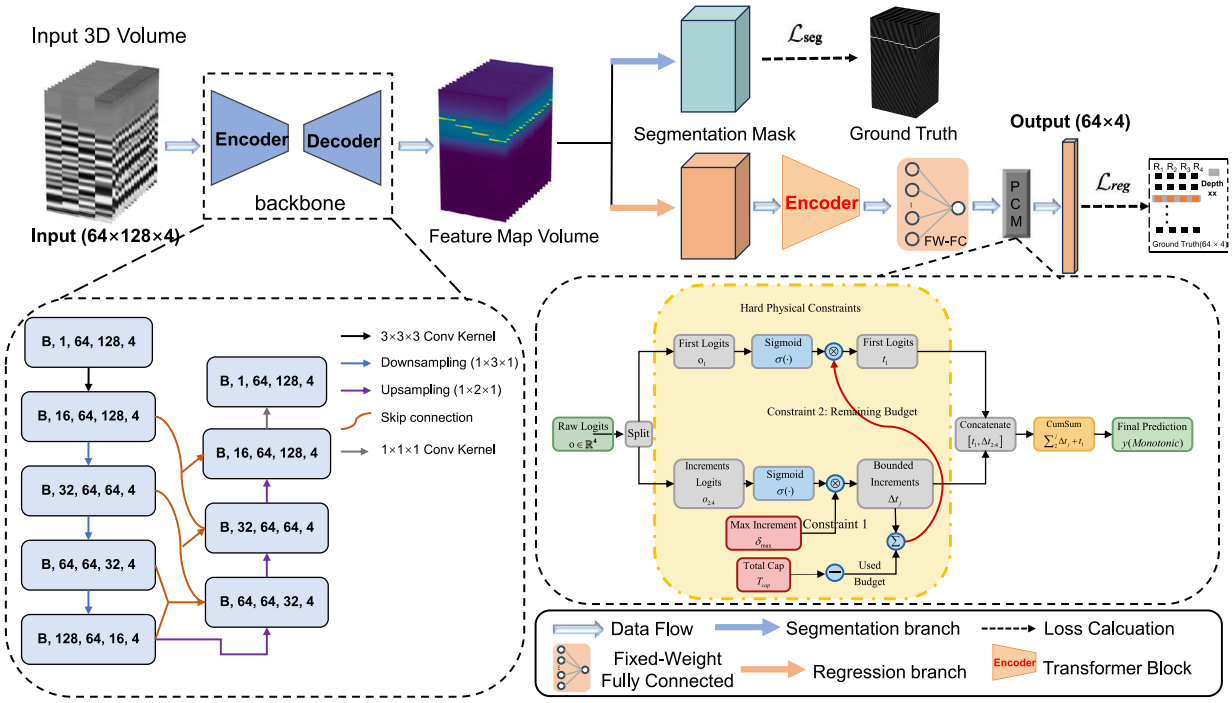


Fig. 7. Architecture of the proposed FA-Picker. FA-Picker consists of a 3D U-Net backbone and two parallel branches: a segmentation branch supervised by mask labels and a regression branch integrating a Transformer with hard physical constraints. Intermediate features extracted by the U-Net facilitate segmentation and assist the regression branch in initial coarse positioning, while hard physical constraints enforce the physical plausibility of the final precise arrival time estimates. In the Hard Physical Constraints module, ⊗ denotes the clamp(·) operation, which enforces the predefined upper constraint boundary.

arrival window, generating feature maps that encode the effective temporal range of the first arrival. These feature maps are explicitly fed into the regression branch as structured contextual guidance, thereby constraining the regression process to a physically meaningful search space [31]. Through this task-level coupling, the network jointly leverages the robustness of segmentation and the precision of regression.

4.2. Backbone 3D U-Net module

We employ a 3D U-Net as the backbone to extract the coarse localization of first-arrival waveforms. A critical advantage of this design lies in the application of 3D convolutional kernels. Unlike 2D approaches, these kernels naturally capture multi-dimensional dependencies [26], simultaneously integrating temporal waveform characteristics, inter-channel moveout patterns, and inter-depth geological continuity.

4.3. Strided convolution

Standard U-Net performs symmetric downsampling across all dimensions, which is suboptimal for array acoustic logging data with narrow depth ($D = 64$) and channel ($C = 4$) dimensions. Downsampling along these axes may introduce padding artifacts and destroy physically meaningful continuity. To address this issue, we apply an anisotropic strided convolution [37] before the 3D convolutional blocks:

$$\mathbf{F}_1 = \mathrm{Conv}_{1\times3\times1}(\mathbf{X}) \tag{9}$$

with padding $(0, 1, 0)$ and stride $(1, 2, 1)$. This pseudo-1D operation performs downsampling only along the time dimension, preserving full resolution in depth and channel directions. Subsequently, standard 3D convolutions are used to learn joint anisotropic representations without loss of spatial structure.

4.4. Self attention

Self-attention provides a learnable mechanism for modeling long-range dependencies in time-series signals through global interactions among sequence elements [29]. In the regression branch, a Transformer encoder is employed to capture temporal correlations in the acoustic waveforms by explicitly modeling global time-to-time relationships. The encoder consists of eight stacked Transformer blocks, which progressively refine the representations without excessive model complexity. Together with multi-head self-attention, this design allows the network to learn anisotropic dependencies across depths, receiver channels, and arrival times.

4.5. Fixed parameters FCL

In standard architectures, the terminal linear layer typically contains learnable weights. In our model, the Transformer encoder outputs a vector $\mathbf{h} \in \mathbb{R}^T$ (here $T = 128$) for each waveform. To obtain scalar arrival time, we apply a fixed linear layer:

$$\hat{t}_{\mathrm{arr}} = \mathbf{w}^{\top}\mathbf{h}, \ \mathbf{w} = \frac{1}{T}\mathbf{1} \tag{10}$$

where $\mathbf{1} \in \mathbb{R}^T$ is an all-ones vector.

This design serves as a deterministic dimensionality reduction. By freezing the weights, the final layer does not introduce bias from uneven arrival-time distributions in the training set. It also prevents shortcut learning: the model cannot rely on the linear layer to memorize dataset statistics, and is forced to extract precise temporal features through the Transformer encoder.

4.6. Hard physical constraints module (PCM)

At each depth sample, the network outputs a four-dimensional vector $\mathbf{o} \in \mathbb{R}^4$. Instead of predicting four arrival times directly, we use a constrained method to ensure the results are physically valid.

First, the outputs o_2, o_3, o_4 are converted into time increments Δt_m using $\Delta t_m = \mathrm{clamp}(\delta_{\max}, \sigma(o_m))$. Here, $\sigma(\cdot)$ is the sigmoid function, and $\mathrm{clamp}(\cdot)$ limits each increment to a maximum value $\delta_{\max}$. In this study, $\delta_{\max}$ is set to 0.095 in the normalized arrival-time domain, corresponding to approximately 12 sampling points, or 96 μs at a sampling interval of 8 μs. Given the inter-receiver spacing s_r, this corresponds to a lower formation-velocity bound of $v_{\min} = s_r/(96\ \mu s)$. The same $\delta_{\max}$ is used in Algorithm 1, ensuring consistency between the feasible model output space and the training-label distribution.

Next, we calculate the remaining time budget B by subtracting the sum of these increments from the total time limit T_{cap} ($B = T_{\mathrm{cap}} - \sum_{m=2}^{4} \Delta t_m$). Since the arrival-time labels are normalized to $[0, 1]$, T_{cap} is set to 1 to ensure that the final predicted arrival time remains within the normalized temporal window. Then, the arrival time of the first receiver (t_1) is calculated from o_1, but it is strictly limited by this budget: $t_1 = \mathrm{clamp}(B, \sigma(o_1))$. This step ensures that the total time for all receivers does not exceed T_{cap}. Finally, the absolute arrival times are obtained by summing these values to guarantee a strictly increasing order:

$$t_i = \begin{cases} t_1, & i = 1 \\ t_1 + \sum_{j=2}^{i} \Delta t_j, & i = 2, 3, 4 \end{cases} \tag{11}$$

where t_i denotes the predicted arrival time for the ith receiver. According to Eq. (11), the four arrival times are constructed in an autoregressive manner, and the non-negative increments Δt_j explicitly guarantee that $t_1 \le t_2 \le t_3 \le t_4$. Therefore, the four predicted arrival-time results are always monotonically non-decreasing.

4.7. Loss function formulation

To ensure the predicted first-arrival times are not only accurate but also physically consistent across array dimensions, we design a composite loss function. This function integrates a segmentation loss with dynamic focusing, gradient similarity constraint, and a physics-informed delay regression loss. The total loss L_{total} is defined as:

$$L_{\mathrm{total}} = \lambda_{\mathrm{seg}} L_{\mathrm{seg}} + \lambda_{\mathrm{grad}} L_{\mathrm{grad}} + \lambda_{\mathrm{reg}} L_{\mathrm{reg}} \tag{12}$$

where λ_{seg}, λ_{grad}, and λ_{reg} are balancing coefficients set to 8.0, 3.0, and 10.0, respectively, to balance the magnitude of gradients between tasks.

4.7.1. Dynamic focal loss for segmentation

Due to the extreme imbalance between first-arrival pixels (positive samples) and background noise, we employ a Dynamic Focal Loss (L_{seg}). Unlike standard focal loss with fixed hyperparameters, we dynamically adjust the focusing parameter γ based on the prediction difficulty [38]. Let p be the predicted probability and $y \in \{0, 1\}$ be the ground truth. The loss is calculated as:

$$L_{\mathrm{seg}} = -\alpha (1 - p_t)^{\gamma(\bar{p}_t)} \log(p_t) \tag{13}$$

where $p_t = p$ if $y = 1$ otherwise $1 - p$, and $\alpha = 0.75$ is a weighting factor. The focusing parameter γ scales linearly with the batch-wise prediction difficulty:

$$\gamma(\bar{p}_t) = \gamma_{\min} + (\gamma_{\max} - \gamma_{\min}) \cdot (1 - \bar{p}_t) \tag{14}$$

where $\bar{p}_t$ represents the mean prediction confidence of the batch. We set $\gamma_{\min} = 2.0$ and $\gamma_{\max} = 5.0$ to intensify training on hard examples as the model converges.

4.7.2. Gradient similarity soft constraint

To encourage the feature representations in the output to vary smoothly across three dimensions, we introduce a gradient similarity loss (L_{grad}), which enforces smoothness in the probability map by minimizing the Mean Squared Error (MSE) between the gradients of the prediction and the target along the time (t), depth (d), and receiver (r) dimensions.:

$$L_{\mathrm{grad}} = \sum_{k \in \{t,d,r\}} w_k \cdot \mathrm{MSE}(\nabla_k p, \nabla_k y) \tag{15}$$

where ∇_k denotes the first-order difference operator along dimension k. Based on the data characteristics, we assign higher weights to time and depth continuity ($w_t = 1.5, w_d = 1.5$) compared to the receiver dimension ($w_r = 0.5$).

Algorithm 2: Physics-constrained optimization workflow of FA-Picker

Input: Waveform volume X; segmentation label Y^{seg}; arrival-time label Y^{arr}.
Output: Physically constrained first-arrival prediction $\hat{Y}^{arr}$.

for *each training iteration* **do**
- Extract spatiotemporal representations from X using the 3D U-Net backbone and generate the auxiliary coarse first-arrival map $\hat{Y}^{seg}$;
- Calculate the segmentation-branch losses, including L_{seg} in Eq. (13) and L_{grad} in Eq. (15);
- Introduce $\hat{Y}^{seg}$ into the regression branch as first-arrival guidance;
- Refine the arrival-sensitive representations using the Transformer Encoder;
- Generate raw logits through the fixed-parameter fully connected layer;
- Enforce the physical constraint in Eq. (11) using the PCM to obtain $\hat{Y}^{arr}$;
- Calculate the arrival-time regression loss in Eq. (16);
- Optimize the trainable parameters by minimizing the total loss in Eq. (12);

return $\hat{Y}^{arr}$;

4.7.3. Physics-constrained regression loss

The regression loss L_{reg} is designed to balance element-wise picking accuracy with spatio-temporal physical consistency. As shown in Eq. (16), it is a weighted sum of three components, where the balancing weights are set to $w_{\mathrm{MAE}} = 1.0$, $w_{\mathrm{cons}} = 0.5$, and $w_{\mathrm{smo}} = 0.2$:

$$L_{\mathrm{reg}} = w_{\mathrm{MAE}} L_{\mathrm{MAE}}(\hat{d}, d) + w_{\mathrm{cons}} L_{\mathrm{cons}} + w_{\mathrm{smo}} L_{\mathrm{smo}} \tag{16}$$

the first term L_{MAE} utilizes the Mean Absolute Error to provide robust supervision for basic picking accuracy.

To incorporate physical constraints, we introduce the consistency loss L_{cons} and the smoothness loss L_{smo}. L_{cons} constrains the first-order derivatives in both spatial dimensions:

$$L_{\mathrm{cons}} = \mathrm{MSE}(\nabla_d \hat{d}, \nabla_d d) + \mathrm{MSE}(\nabla_r \hat{d}, \nabla_r d) \tag{17}$$

Physically, explicit constraints on ∇_d and ∇_r enforce smooth arrival-time variations along the borehole depth and across the receiver array, respectively.

Finally, to encourage the linear moveout property of the array, L_{smo} imposes a second-order smoothness constraint along the receiver dimension:

$$L_{\mathrm{smo}} = \mathrm{MSE}(\nabla_r^2 \hat{d}, \mathbf{0}) \tag{18}$$

where $\nabla_r^2 \hat{d}$ denotes the second-order finite difference of the predicted arrival times along the receiver dimension, and $\mathbf{0}$ is a zero tensor with the same size as $\nabla_r^2 \hat{d}$. For a 4-receiver array, this term encourages the three adjacent inter-receiver arrival-time differences to be close to each other, thereby suppressing high-frequency jitter and maintaining consistency with the inherent moveout trends.

Table 2
Number of depth points and SNR values of the eight datasets.

Dataset	W1	W2	W3	W4	W5	W6	W7	W8
Depth points	29 321	25 819	2805	8369	13 376	11 231	16 829	12 092
SNR (dB)	0.386	0.847	0.087	0.063	0.139	−0.280	−0.132	−1.204

4.8. Optimization workflow

The complete optimization procedure of FA-Picker is summarized in Algorithm 2. The segmentation output is used only as auxiliary guidance, whereas the physically constrained regression output $\hat{Y}^{arr}$ is adopted as the final first-arrival picking result.

5. Experiments

5.1. Settings

5.1.1. Dataset description

To evaluate the performance and generalization capability of the proposed model, eight wells with different noise levels were used for testing. Table 2 lists the number of depth points and the SNR value of each dataset. Based on these SNR values, W1–W2 were classified as high-SNR datasets, W3–W5 as medium-SNR datasets, and W6–W8 as low-SNR datasets. The wells are distributed across different regions and formation development patterns, enabling a comprehensive assessment of the model under varying data quality conditions.

5.1.2. Preprocessing

Sliding window. A sliding window of length 64 is designed and moved along the depth direction of each well to generate prediction samples. To ensure prediction stability, an overlap of four depth samples is employed between adjacent windows, corresponding to a sliding step size of 60. For the overlapping regions, the final prediction is obtained by averaging the results from the two adjacent windows.

Normalization. For data preprocessing, only mean–variance normalization is applied to the raw waveforms. Notably, the normalization is performed independently for each waveform trace, which preserves the intrinsic waveform characteristics of individual receivers.

5.1.3. Evaluation metrics

For array acoustic logging, we use the average moveout at each depth point as the evaluated target. It is calculated by averaging the three inter-receiver arrival-time differences among the four receiver channels. Model performance is evaluated by comparing the predicted and ground-truth average moveouts using four metrics: accuracy (Acc), root mean square error (RMSE), mean Pearson correlation (M-Pearson), and global Pearson correlation (G-Pearson). A prediction is regarded as accurate when the absolute error between the predicted and ground-truth average moveout is below 8 μs.

M-Pearson is introduced to evaluate local slowness-pattern consistency. Since each training and inference sample contains 64 consecutive depth points, the Pearson correlation is computed within each 64-point segment and then averaged over all segments. This metric reflects whether the model preserves local geological variations. For each well, consecutive average moveout values from $L = 64$ depth points are grouped into a one-dimensional sequence, and M-Pearson is defined as follows:

$$\text{M-Pearson} = \frac{1}{M}\sum_{i=1}^{M}\frac{\sum_{j=1}^{L}\left(y_{i,j}-\bar{y}_i\right)\left(\hat{y}_{i,j}-\bar{\hat{y}}_i\right)}{\sqrt{\sum_{j=1}^{L}\left(y_{i,j}-\bar{y}_i\right)^2}\sqrt{\sum_{j=1}^{L}\left(\hat{y}_{i,j}-\bar{\hat{y}}_i\right)^2}}, \tag{19}$$

where M denotes the number of segments, $L = 64$ is the segment length, y and $\hat{y}$ represent the ground-truth and predicted average moveouts, respectively, and $\bar{y}_i$ and $\bar{\hat{y}}_i$ are the corresponding segment means.

For completeness, G-Pearson is also reported as the standard Pearson correlation coefficient computed over all depth points of each well. Thus, M-Pearson measures local 64-point consistency, whereas G-Pearson measures global trend consistency.

Table 3
STC parameter settings.

Setting	Window	Shift	Semblance	Energy
STC-A	8	5	0.55	2.0
STC-B	4	7	0.55	2.5
STC-C	3	7	0.50	2.5

Table 4
Overall comparison of first-arrival picking methods.

Method	Type	Acc (%) ↑	RMSE (μs) ↓	M-Pearson ↑
STC-A	Traditional	93.24	9.41	0.28
STC-B	Traditional	89.63	11.00	0.30
STC-C	Traditional	86.63	12.69	0.26
MS-Picking	Deep learning	81.80	33.16	0.41
2D U-Net	Deep learning	96.14	3.68	0.74
DAPhaseNet	Deep learning	91.80	16.57	0.55
FA-Picker	**Proposed**	**98.29**	**2.31**	**0.87**

Note: Bold values indicate the best performance in each category.

5.2. Comparison and ablation study

5.2.1. Baseline comparison

To comprehensively evaluate the effectiveness of the proposed FA-Picker, both traditional and deep learning-based first-arrival picking methods are included for comparison. For the traditional baseline, since STC does not involve trainable parameters, its performance mainly depends on the selected processing configuration. Therefore, three representative parameter settings, denoted as STC-A, STC-B, and STC-C, are tested to provide a more robust comparison with learning-based methods. The detailed parameter settings are listed in Table 3.

For the deep learning baselines, MS-Picking [15], 2D U-Net [39], and DAPhaseNet [13] are included. These models represent typical 2D or depth-independent learning strategies, in which each depth point is processed individually. In contrast, FA-Picker takes a 3D data volume as input to explicitly exploit inter-depth continuity. To ensure a fair comparison, the training samples for all 2D deep learning baselines are derived from the same 3D training dataset used by FA-Picker. Specifically, the 32^{nd} depth point, the center depth point, is extracted from each 3D sample to construct the corresponding 2D input. Thus, all deep learning baselines are trained with the same data source and label distribution, while differing in their input formulation and model architecture.

All deep learning models are trained using the same optimization settings. The training and validation data are split at a ratio of 8:2, the AdamW optimizer is used, and the initial learning rate is set to 0.001. For overall evaluation, the predictions from all test wells are aggregated, and the Acc, RMSE, and M-Pearson metrics are calculated based on the average moveout over the complete test set.

As shown in Table 4, FA-Picker achieves the best overall performance among all compared methods. Compared with the strongest deep learning baseline, 2D U-Net, FA-Picker improves the Acc, reduces the RMSE, and obtains higher correlation metrics. Among the traditional methods, STC-A obtains the best performance, but its RMSE and correlation scores remain clearly inferior to those of FA-Picker. These results demonstrate the advantage of combining 3D inter-depth continuity modeling with physics-constrained prediction under complex field conditions.

Since the above evaluation is based on the average moveout, it mainly reflects the overall inter-receiver arrival-time trend and may not

Table 5
Receiver-wise first-arrival RMSE over all eight test wells (μs).

Method	Receiver 1	Receiver 2	Receiver 3	Receiver 4
STC-A	90.50	92.80	95.19	98.09
STC-B	83.34	87.02	91.66	96.96
STC-C	86.21	90.85	96.64	103.37
MS-Picking	98.36	100.81	92.56	86.88
2D U-Net	36.85	38.20	39.52	40.87
DAPhaseNet	57.85	70.25	73.18	74.89
FA-Picker	**11.01**	**11.51**	**12.22**	**13.01**

fully reveal the prediction accuracy of each individual receiver. In particular, for a four-receiver array, the average moveout is computed from the adjacent inter-receiver differences and therefore does not explicitly evaluate the absolute first-arrival error of each receiver channel. To provide a more detailed assessment, we further calculate the receiver-wise first-arrival RMSE over all eight test wells, as reported in Table 5. The results show that FA-Picker achieves substantially lower RMSE than all baseline methods on all four receiver channels, indicating that its improvement is not limited to the moveout-level metric but is also reflected in the direct receiver-wise arrival-time prediction.

5.2.2. Component and 3D modeling ablation

To assess the contributions of key components in FA-Picker, ablation studies are performed under different SNR conditions, as reported in Table 6. The evaluated variants include FA-w/o-Trans, FA-2DVariant, FA-w/o-PCM, and the FA-Picker. FA-w/o-Trans removes the Transformer encoder to evaluate the contribution of long-range depth-wise dependency modeling. FA-w/o-PCM removes the Physical Constraint Module to assess the effect of physics-constrained prediction. To further isolate the contribution of 3D spatiotemporal modeling, we add a controlled variant named FA-2DVariant. This variant preserves the main prediction framework of FA-Picker but adapts the input and backbone to a 2D setting.

As shown in Table 6, removing the Transformer encoder causes substantial performance degradation, confirming the importance of long-range depth-wise dependency modeling. The FA-2DVariant also performs consistently worse than FA-Picker, particularly under low-SNR conditions. These results demonstrate that explicit 3D spatiotemporal modeling is essential for improving robustness and maintaining coherent first-arrival trajectories in complex field data. It is worth noting that FA-Picker achieves an RMSE of only 0.84 μs on W_2, which is much smaller than the original sampling interval of 8 μs. In this well, 99.87% of the predictions are within one sampling interval, and 94.50% have zero sampling-point error after both predictions and references are quantized to the nearest integer sampling point. This sub-sampling-level RMSE is obtained because the labels are refined to continuous sub-sample positions through interpolation, and the regression branch predicts continuous arrival times rather than discrete sample indices.

Removing the PCM also generally weakens the model. FA-Picker achieves higher Acc and lower RMSE than FA-w/o-PCM in seven of the eight wells, indicating that the physical constraint improves prediction robustness and suppresses physically inconsistent arrivals in most cases. The only exception occurs in the extremely low-SNR well W_8, where FA-w/o-PCM obtains higher Acc and M-Pearson, while both models have the same RMSE of 4.60 μs. Representative examples in Fig. 8 show that the first arrivals in W_8 are weak, discontinuous, and close to the right boundary of the 128-sample temporal window. In this boundary-adjacent condition, the hard autoregressive constraint of the PCM narrows the feasible prediction space because the arrivals must satisfy $t_1 \le t_2 \le t_3 \le t_4$ and $t_4 \le T_{\mathrm{cap}}$. Small errors in the estimated inter-receiver increments can therefore alter the remaining time budget for t_1 and push the prediction toward a boundary-saturated solution, resulting in local depth-wise jumps.

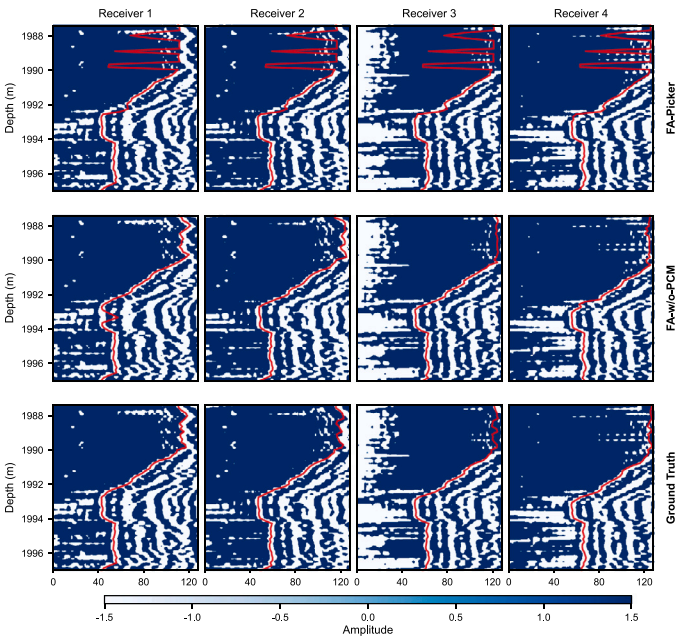


Fig. 8. Representative boundary-adjacent first-arrival picking case in the extremely low-SNR well W_8. From top to bottom: FA-Picker, FA-w/o-PCM, and ground truth.

As a preliminary mitigation analysis for the W_8 boundary case, we further evaluated the effect of sliding-window overlap by varying the overlap from 0 to 32 samples while keeping the trained model unchanged. The resulting performance variations are minor, with Acc ranging from 92.71% to 93.03% and RMSE ranging from 4.48 to 4.60 μs. This limited sensitivity suggests that simply increasing the window overlap does not substantially alleviate the observed degradation, and that insufficient overlap is unlikely to be its primary cause. Enriching the training data with more boundary-adjacent and rapidly varying samples may therefore be a more promising mitigation strategy.

5.2.3. Sensitivity analysis of $\delta_{\max}$

In addition to the component ablation, we further conduct a sensitivity analysis on the maximum increment parameter $\delta_{\max}$ in the PCM. The baseline value is set to $\delta_{\max} = 0.095$ in the normalized arrival-time domain. Two additional settings with ±5% perturbations are tested to evaluate the robustness of this parameter. As shown in Table 7, the baseline setting achieves the lowest RMSE, suggesting that $\delta_{\max} = 0.095$ provides a suitable balance between physical feasibility and prediction flexibility. These results support the robustness of FA-Picker to moderate changes in the maximum inter-receiver increment constraint.

5.3. Evaluation of real-time performance

FA-Picker contains 2.665 M parameters and requires approximately 17.998 GFLOPs for a input volume. To verify the real-time inference performance of FA-Picker in field deployment scenarios, the proposed algorithm was implemented on a C++ inference platform and evaluated under realistic hardware constraints. Considering the limited computational resources typically available in logging environments, experiments were conducted on a low-power laptop equipped with a 4-core, 8-thread CPU operating at 1.00 GHz. To closely emulate on-site logging conditions, a sequential depth-by-depth data acquisition process was simulated, where waveform data are continuously collected and processed in real time. The average inference latency was

Table 6
Performance comparison of ablation models under different SNR conditions.

Model	Metric	High SNR		Medium SNR			Low SNR		
		W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
FA-w/o-Trans	Acc (%)	82.06	75.00	70.09	98.41	82.92	57.64	39.85	35.96
	RMSE (μs)	6.30	6.57	7.70	2.47	6.54	7.79	9.14	10.12
	M-Pearson	0.76	0.51	0.63	0.55	0.67	0.32	0.53	0.50
	G-Pearson	0.90	0.58	0.77	0.89	0.77	0.35	0.71	0.69
FA-2DVariant	Acc (%)	98.53	97.53	77.54	96.53	97.98	82.84	78.35	76.74
	RMSE (μs)	2.72	3.02	8.82	3.95	2.71	7.14	8.74	9.43
	M-Pearson	0.82	0.86	0.53	0.65	0.77	0.38	0.34	0.35
	G-Pearson	0.90	0.88	0.60	0.72	0.86	0.47	0.43	0.53
FA-w/o-PCM	Acc (%)	99.60	99.81	94.58	98.72	99.59	94.23	96.77	**96.49**
	RMSE (μs)	1.57	1.02	7.59	2.35	1.18	6.36	5.40	4.60
	M-Pearson	0.90	0.97	0.81	0.92	0.94	0.66	0.79	**0.80**
	G-Pearson	0.96	0.98	0.71	0.91	0.97	0.58	0.65	0.77
FA-Picker	Acc (%)	**99.75**	**99.87**	**97.86**	**98.96**	**99.74**	**96.84**	**96.89**	92.71
	RMSE (μs)	**1.31**	**0.84**	**2.59**	**1.75**	**1.17**	**3.31**	**2.76**	4.60
	M-Pearson	0.90	**0.98**	**0.90**	0.92	0.94	0.66	**0.82**	0.75
	G-Pearson	**0.97**	**0.99**	**0.94**	**0.95**	0.97	**0.81**	**0.86**	**0.78**

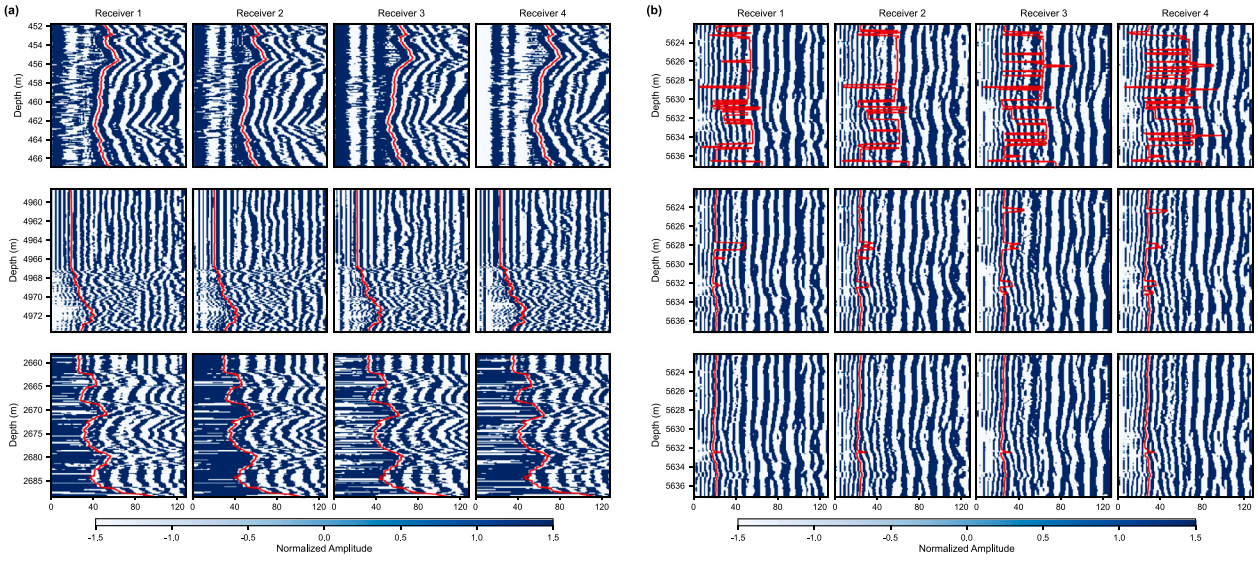


Fig. 9. Visualization of prediction results on field data. (a) Performance of the proposed FA-Picker on three complex well sections. (b) Comparison of three different models on the same well section. From top to bottom, the rows display results from the 2D U-Net baseline, FA-w/o-PCM, and FA-Picker.

Table 7
Sensitivity analysis of the maximum increment parameter δ_{max}.

δ_{max}	Variation	Acc (%)	RMSE (μs)
0.09025	−5%	98.1442	2.8601
0.09500	**Baseline**	**98.2894**	**2.3081**
0.09975	+5%	98.2410	2.9283

calculated by dividing the total inference time by the number of output depth points, resulting in approximately 18 ms per depth point, which fully satisfies the real-time requirements.

5.4. Cases analys

5.4.1. Generalization performance test

The generalization performance was evaluated on the field test set comprising 119,842 depth samples from eight wells. Specifically, the trained 2D U-Net baseline, FA-w/o-PCM, and FA-Picker were applied to this test set under the same inference protocol. Several representative well sections were selected from the prediction results for visualization, including three challenging cases for FA-Picker and one common well section for comparing the three models. As shown in Fig. 9(a), FA-Picker reliably identifies weak first arrivals under strong noise and tracks arrival-time variations caused by stratigraphic changes, producing smooth and geophysically consistent predictions along the depth direction. Beyond this local visualization, Fig. 10 presents a sample-wise comparison over the entire field test set. The overall R^2 of 0.9169 demonstrates strong large-scale agreement between the predicted and manually picked first arrivals.

5.4.2. Importance of physical constraints

Fig. 9(b) demonstrates how physical constraints improve prediction consistency. Unconstrained models can produce invalid results, such as later arrivals at the first receiver than at subsequent ones. In contrast, FA-Picker enforces inter-receiver consistency, maintaining alignment to the same wave phase even when minor deviations occur. Unlike 2D U-Net or unconstrained models, which show depth-wise jumps and discontinuities, our model incorporates the depth dimension to ensure smooth, continuous predictions, demonstrating that combining spatial–temporal continuity with physical constraints is essential for robust first-arrival picking.

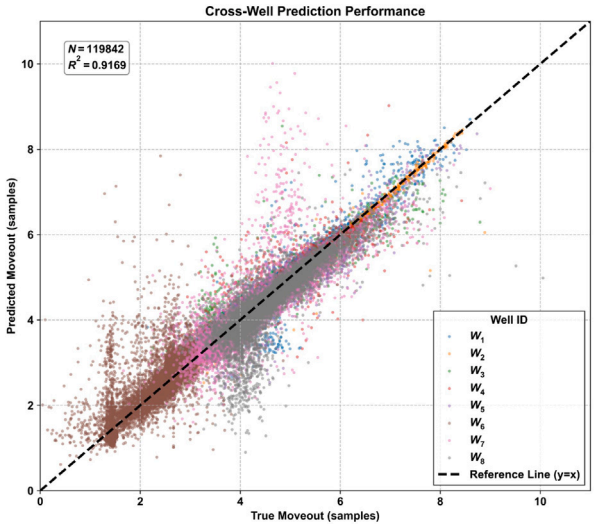


Fig. 10. Performance evaluation and cross-validation of first-arrival moveout prediction across eight unseen wells. The scatter plot correlates ground-truth labels with FA-Picker predictions, where different colors represent distinct geological environments (W_1–W_8). The high coefficient of determination ($R^2 = 0.9169$) across $N = 119842$ samples, relative to the $y = x$ reference (dashed line), demonstrates the robust generalization of the physics-constrained model in diverse formation conditions.

6. Conclusion

This paper proposes FA-Picker, a physically constrained deep learning model for real-time first-arrival picking in array acoustic logging. To address the challenge of high-quality label scarcity in structured time-series sensing, we developed a comprehensive synthetic data pipeline that embeds domain knowledge to simulate diverse geological scenarios. By incorporating a hard physical constraint module, we successfully trained a reliable supervised model. Extensive experiments on real-well datasets demonstrate that FA-Picker ensures geophysical consistency while maintaining superior accuracy and stability.

This research provides a viable solution for pattern recognition in highly structured sensing data where manual annotation is costly or unavailable. The core contribution lies in the effective integration of domain-informed data synthesis and physical constraints, offering a robust framework for building reliable models in expert-dependent fields. This "physics-informed + data-driven" approach serves as a valuable reference for other industrial sensing tasks, proving that specialized supervised models can achieve high reliability and real-time performance in specialized vertical domains.

Despite its effectiveness, FA-Picker is currently limited by its fixed 4-channel input and the assumption of a specific monotonic arrival sequence. In practice, varying logging instruments or acquisition schemes may yield different receiver counts or even reversed arrival sequences, necessitating additional data adaptation. In addition, the limited temporal extent of each waveform may reduce robustness when the true first arrivals are close to the temporal boundary. Such boundary-adjacent cases are relatively rare in the training data, and the hard physical constraints may occasionally amplify local uncertainty under extremely low-SNR conditions. Furthermore, the extreme complexity of real-world borehole signals continues to challenge the generalization of supervised learning. Future work will focus on more advanced data synthesis strategies, boundary-aware training samples, adaptive input handling, and robust abnormal-data detection, as well as the development of large-scale foundation models capable of processing multi-dimensional and multi-modal logging signals. Such advancements aim to transition array acoustic logging from a process reliant on expert experience to a fully automated, intelligent paradigm.

CRediT authorship contribution statement

Long Wang: Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Xinming Wu:** Writing – review & editing, Visualization, Validation, Methodology, Funding acquisition, Conceptualization. **Zhifeng Sun:** Visualization, Methodology, Data curation. **Yimin Dou:** Writing – review & editing, Validation, Methodology.

Declaration of generative AI use

During the preparation of this manuscript, the authors used Gemini to improve language clarity and readability. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Young Scientists Fund, Category C; Grant No. 42504108), National Postdoctoral Program for Innovative Talents (Grant No. BX20250436), the National Key Research and Development Program of China (Grant No. 2024YFC3012803-05), and the Natural Science Foundation of Anhui Province (Youth Project; Grant No. 2408085QD127).

Data availability

Data will be made available on request.

References

[1] X.-M. Tang, C.H.A. Cheng, Quantitative Borehole Acoustic Methods, vol. 24, Elsevier, Amsterdam, The Netherlands, 2004.
[2] D.V. Ellis, J.M. Singer, Well Logging for Earth Scientists, second ed., Springer, Dordrecht, The Netherlands, 2007.
[3] J.L Mari, C. Vergniault, F. Coppens, Acoustic logging, in: J.-L. Mari, C. Vergniault (Eds.), Acoustic Wave Propagation in Heterogeneous Media, EDP Sciences, Les Ulis, France, 2020, pp. 77–102.
[4] L. Kuang, et al., Application and development trend of artificial intelligence in petroleum exploration and development, Pet. Explor. Dev. 48 (1) (2021) 1–14.
[5] M.E. Willis, M.N. Toksöz, Automatic P and S velocity determination from full waveform digital acoustic logs, Geophysics 48 (12) (1983) 1631–1644.
[6] C.V. Kimball, T.L. Marzetta, Semblance processing of borehole acoustic array data, Geophysics 49 (3) (1984) 274–281.
[7] J. Akram, D.W. Eaton, A review and appraisal of arrival-time picking methods for downhole microseismic data, Geophysics 81 (2) (2016) KS71–KS91.
[8] J. Wong, L. Han, J. Bancroft, R. Stewart, Automatic time-picking of first arrivals on noisy microseismic data, in: Proc. CSEG Conf., Calgary, AB, Canada, 2009, pp. 1–4.
[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[10] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognit. 120 (2021) 108102.
[11] Z. Li, L. Zhu, T. Officer, F. Shi, T. Yu, Y. Wang, A machine-learning-based method of detecting and picking the first P-wave arrivals of acoustic emission events in laboratory experiments, Geophys. J. Int. 230 (3) (2022) 1818–1823.
[12] T.Y. Guo, T. Vanorio, J. Ding, A deep-learning p-wave arrival picker for laboratory acoustic emissions: Model training and its performance, Rock Mech. Rock Eng. 58 (3) (2025) 3073–3091.
[13] L. Zhu, H. Li, L. Ma, J. Li, B. Wu, J. Gao, Intelligent identification first arrivals of acoustic logging curves using dual attention PhaseNet, IEEE Geosci. Remote. Sens. Lett. 21 (2024) 1–5.
[14] J. Zhang, G. Sheng, First arrival picking of microseismic signals based on nested U-net and Wasserstein generative adversarial network, J. Pet. Sci. Eng. 195 (2020) 107527.

[15] G. Sheng, Z. Zhang, X. Tang, K. Xie, A first arrival picking method of microseismic signals based on semi-supervised learning using FreeMatch and MS-picking, Comput. Geosci. 196 (2025) 105844.
[16] X. Si, X. Wu, Z. Li, S. Wang, J. Zhu, An all-in-one seismic phase picking, location, and association network for multi-task multi-station earthquake monitoring, Commun. Earth Env. 5 (2024) 22.
[17] X. Li, Y. Wang, J. Xing, Y. Wang, Causal graph inference with adaptive dynamic structure learning for mechanism-oriented fault diagnosis in dynamic industrial systems, Reliab. Eng. Syst. Saf. 266 (2026) 111865.
[18] X. Li, J. Zhang, K. Zhu, Q. Hu, L. Sun, J. Gu, W. Lyu, Y. Wang, Interpretable and physics-informed graph neural modeling for intelligent damage localization in CFRP composites, Compos. Part B. 315 (2026) 113469.
[19] J.H. Doveton, Geologic Log Analysis Using Computer Methods, American Association of Petroleum Geologists, Tulsa, OK, USA, 1994.
[20] C. Guo, T. Zhu, Y. Gao, S. Wu, J. Sun, AEnet: Automatic picking of P-wave first arrivals using deep learning, IEEE Trans. Geosci. Remote Sens. 59 (6) (2021) 5293–5303.
[21] W. Zhu, G.C. Beroza, PhaseNet: a deep-neural-network-based seismic arrival-time picking method, Geophys. J. Int. 216 (1) (2019) 261–273.
[22] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, 2017, pp. 843–852.
[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[24] S. Yu, J. Ma, Deep learning for geophysics: Current and future trends, Rev. Geophys. 59 (3) (2021) e2021RG000742.
[25] H.M. Ahmad, D. Morle, A. Rahimi, LayerMix: Enhanced data augmentation for robust deep learning, Pattern Recognit. 172 (2026) 112332.
[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, Santiago, Chile, 2015, pp. 4489–4497.
[27] Z.E. Ross, M.-A. Meier, E. Hauksson, T.H. Heaton, Generalized seismic phase detection with deep learning, Bull. Seismol. Soc. Am. 108 (5A) (2018) 2894–2901.
[28] Z. Wang, J. Yao, M. Xu, M. Jiang, J. Su, Transformer-based network with temporal depthwise convolutions for sEMG recognition, Pattern Recognit. 145 (2024) 109967.
[29] A. Vaswani, et al., Attention is all you need, NIPS, in: Adv. Neural Inf. Process. Syst., vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.
[30] Y. Yu, R. Ma, Z. Ma, Robformer: A robust decomposition transformer for long-term time series forecasting, Pattern Recognit. 153 (2024) 110552.
[31] X. Sun, B. Xiao, F. Wei, L. Shuang, Y. Wei, Integral human pose regression, in: Proc. Eur. Conf. Comput. Vis., ECCV, Munich, Germany, 2018, pp. 529–545.
[32] S. Mu, B. Liu, S. Jin, C. Lien, Y. Zhang, W. Xie, FCT-stock: A fusion CNN-transformer model for anomaly detection in stock volatility, Pattern Recognit. 177 (2026) 113260.
[33] B. Boashash, Estimating and interpreting the instantaneous frequency of a signal—Part I: Fundamentals, Proc. IEEE 80 (4) (1992) 520–538.
[34] C.H. Cheng, F.L. Paillet, W.D. Pennington, Acoustic waveform logging—Advances in theory and application, Log Anal. 32 (3) (1992) 239–258.
[35] Xiao-Ming Tang, Zhen Li, Chuang Hei, Yuan-Da Su, Elastic wave scattering to characterize heterogeneities in the borehole environment, Geophys. J. Int. 205 (1) (2016) 594–603.
[36] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, Venice, Italy, 2017, pp. 2961–2969.
[37] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: Proc. Int. Conf. Learn. Represent., ICLR, San Diego, CA, USA, 2015.
[38] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, NeurIPS, in: Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 20913–20924.
[39] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proc. MICCAI, Cham, Switzerland, 2015, pp. 234–241.