



## OPEN MFU-Net: a multi-scale fusion U-Net for seismic phase picking

Lihua Wu<sup>1,3</sup>, Jiaquan Yan<sup>2,3</sup>, Yanming Zhang<sup>1</sup>, Mingquan Hong<sup>1</sup>, Hongcai Zhang<sup>1</sup> & Shirong Liao<sup>1</sup>✉

Seismic phase picking aims to accurately identify and label the arrival times of different types of seismic waves (e.g., P-waves and S-waves) from waveform data, serving as a fundamental step in seismological research and related applications. Although existing deep learning-based methods have achieved notable accuracy, their architectural designs are often overly complex. In this paper, we propose a Multi-Scale Fusion U-Net architecture (MFU-Net), aiming to achieve significant recognition performance through straightforward enhancements to the traditional U-Net design. Specifically, we design a multi-scale feature fusion module within the skip connections of the U-Net architecture to effectively integrate multi-scale semantic and spatial information. This is followed by the incorporation of a multi-head attention mechanism in the bottleneck layer to enhance the recognition of critical feature regions. Finally, a weighted class-balanced loss is introduced into the loss function to improve the model's ability to identify minority classes. Tests conducted on the seismic dataset provided by the Fujian Earthquake Agency and the open-source STEAD dataset show that, compared to GPD, SegPhase, and SEANet, the proposed MFU-Net achieves improvements of 1.6% and 1.4% in P-wave picking accuracy, and enhancements of 4.1% and 2.7% in S-wave picking accuracy, respectively.

**Keywords** Underwater image enhancement, Disentanglement strategy, Multiscale feature fusion, Underwater optical imaging

Seismic phase picking is fundamental to seismic data analysis, as accurate phase identification aids in earthquake localization and focal mechanism determination<sup>1,2</sup>. With the increasing number of seismic stations and the explosion of recorded signals, automatic picking techniques have replaced traditional manual methods<sup>3</sup>. However, these early automatic picking techniques are constrained by manually-defined thresholds and demonstrate limited robustness in complex environments, as they rely on constructing characteristic functions from waveform attributes (e.g., amplitude and frequency) and applying predefined thresholds for phase detection. For example, STA/LTA<sup>4</sup> and AR-AIC<sup>5</sup> require experts to manually set up parameters, which are sensitive to the detection threshold. The SCC technique utilizes template waveforms to compute cross-correlation coefficients of seismic data, followed by comparator thresholding for precise P-wave arrival time detection<sup>6</sup>. Experimental results reveal that our MD-Net produces more realistic underwater colors and clearer visibility and shows favorable results in both trainable parameters (Params) and running times (Runtimes). The main contributions of this paper can be summarized as follows:

With the rapid advancement of computing technology, deep learning technologies have been increasingly adopted in seismological applications. The PhaseNet<sup>7</sup> is the first to formulate seismic phase picking as a point-wise classification task and uses the U-Net architecture<sup>8</sup> to perform point-by-point predictions on continuous waveforms, which significantly improved the accuracy of P-wave and S-wave picking. The stacked denoising autoencoder (SDAE)<sup>9</sup> suppresses background noise to enhance the clarity and sharpness of phase onsets, followed by a hard-decision thresholding operation for event detection. Li et al.<sup>10</sup> modified the UNet++<sup>11</sup> architecture based on the characteristics of seismic waveform data and phase arrival picking, then trained the model using data from the China Seismic Experimental Site (CSES) and proposed a phase picking network with strong generalization capability, named CSESnet. Ma et al.<sup>12</sup> treated the phase picking problem as a binary image segmentation task and employed an improved pixel-wise convolutional network to automatically predict phase arrival times. Wang et al.<sup>13</sup> developed the PickNet model, a convolutional neural network with multi-branch supervision designed to detect seismic arrival times. Hou et al.<sup>14</sup> proposed a plug-and-play Sequential Attention (SEA) module to extract the sequential characteristics of seismic signals and constructed a fully convolutional neural network with a sequential attention mechanism to detect seismic events and determine phase arrival times. Li et al.<sup>15</sup> integrated attention mechanisms into the U-Net<sup>8</sup> architecture and proposed a deep learning-

<sup>1</sup>Fujian Earthquake Agency, Fuzhou 350003, China. <sup>2</sup>Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, School of Computer and Data Science, Minjiang University, Fuzhou 350121, China. <sup>3</sup>Lihua Wu and Jiaquan Yan contributed equally to this work. ✉email: liaoshirong@fjea.gov.cn

based model called EPick, which is capable of extracting hierarchical deep features, and its decoder fully leverages the multi-scale features extracted by the encoder to achieve accurate phase picking. Chai et al.<sup>16</sup> applied transfer learning to the PhaseNet<sup>7</sup> to handle data with spatial and temporal feature variations, and developed a new phase picking workflow by integrating deep learning with double-difference earthquake imaging. Lan et al.<sup>17</sup> introduced a cross-attention network model optimized for the NVIDIA Jetson Xavier NX platform for seismic phase picking, which leverages a cross-attention mechanism to integrate time-domain and frequency-domain information, thereby enhancing feature recognition accuracy, reducing dependence on central servers in edge computing environments, and enabling real-time seismic monitoring. Zhu et al.<sup>18</sup> proposed a lightweight hybrid network that integrates convolutional neural networks with Transformers. The model employs an improved self-attention mechanism to extract global contextual features, while convolutional layers are used to preserve local information from the input features. This hybrid architecture facilitates effective interaction between global features and local details, enabling more accurate capture of phase arrival times. In addition, Yu et al.<sup>19</sup> analyzed the performance of various deep learning-based phase pickers using the DiTing dataset and provided recommendations for model selection tailored to specific application scenarios. The development of the above-mentioned methods is particularly important for advancing real-time intelligent seismic processing systems<sup>20</sup>. Given the requirements for high detection accuracy and real-time processing capability in seismic phase picking tasks, and considering the demonstrated success of PhaseNet<sup>7</sup>, we contend that the U-Net architecture serves as an appropriate foundational model for this task.

Therefore, we analyze the limitations of the traditional U-Net architecture<sup>8</sup> and optimize its detection performance at minimal computational cost. In the traditional U-Net architecture<sup>8</sup>, skip connections concatenate feature maps from corresponding encoder and decoder layers directly as input to subsequent decoder. This approach results in insufficient fusion of multi-scale semantic and spatial information, thereby limiting the further improvement of the U-Net architecture<sup>8</sup>. Moreover, the bottleneck layer located between the encoder and decoder is responsible for capturing the global contextual information of the input data. But, this layer typically applies standard convolution operations uniformly across all features, making it incapable of distinguishing between critical and non-critical regions. As a result, important seismic phase information may be overwhelmed by background noise. In addition, the seismic phase picking task suffers from extreme class imbalance in the training data. Typically, the vast majority of time points belong to the “noise” class, while true P-wave and S-wave instances account for only a small fraction.

To mitigate the above three problems, we first propose a multi-scale fusion U-Net. Specifically, in the skip connections of the traditional U-Net<sup>8</sup>, we integrate a Multi-scale Feature Fusion Module<sup>21</sup> (MFFM) based on a convolutional weighting mechanism to address the limitation of skip connections in traditional U-Net architectures, which fail to effectively fuse semantic and spatial information. The MFFM<sup>21</sup> module adaptively integrates multi-level features from the encoder and decoder by learning fusion weights, thereby significantly enhancing the accuracy and robustness of seismic phase picking. Meanwhile, considering that the bottleneck layer carries global contextual information, we introduce a Multi-head Attention Mechanism<sup>22</sup> (MAM) at this stage to guide the network's focus toward key regions such as seismic phase signals. This significantly suppresses background noise and redundant information, thereby improving the discriminative ability of the model in phase picking tasks. Finally, given the severe class imbalance in seismic phase picking tasks, we incorporate a class-weighting strategy in the loss function, assigning higher weights to rare classes. This improves the model's sensitivity and recognition capability for minority classes, thereby enhancing its generalization performance in real-world applications.

The main contributions of this paper are summarized as follows:

1. We design a multi-scale feature fusion module (MFFM) to integrate spatial and semantic features from different levels, enhancing the accuracy and robustness of the traditional U-Net in seismic phase picking.
2. We incorporate a multi-head attention mechanism (MAM) in the bottleneck layer to enhance the network's ability to perceive key signal regions.
3. We introduce a class-weighted loss function to mitigate the effects of sample imbalance, increasing the weights of rare classes (P-waves and S-waves) in the loss computation to enhance the network's focus on them.

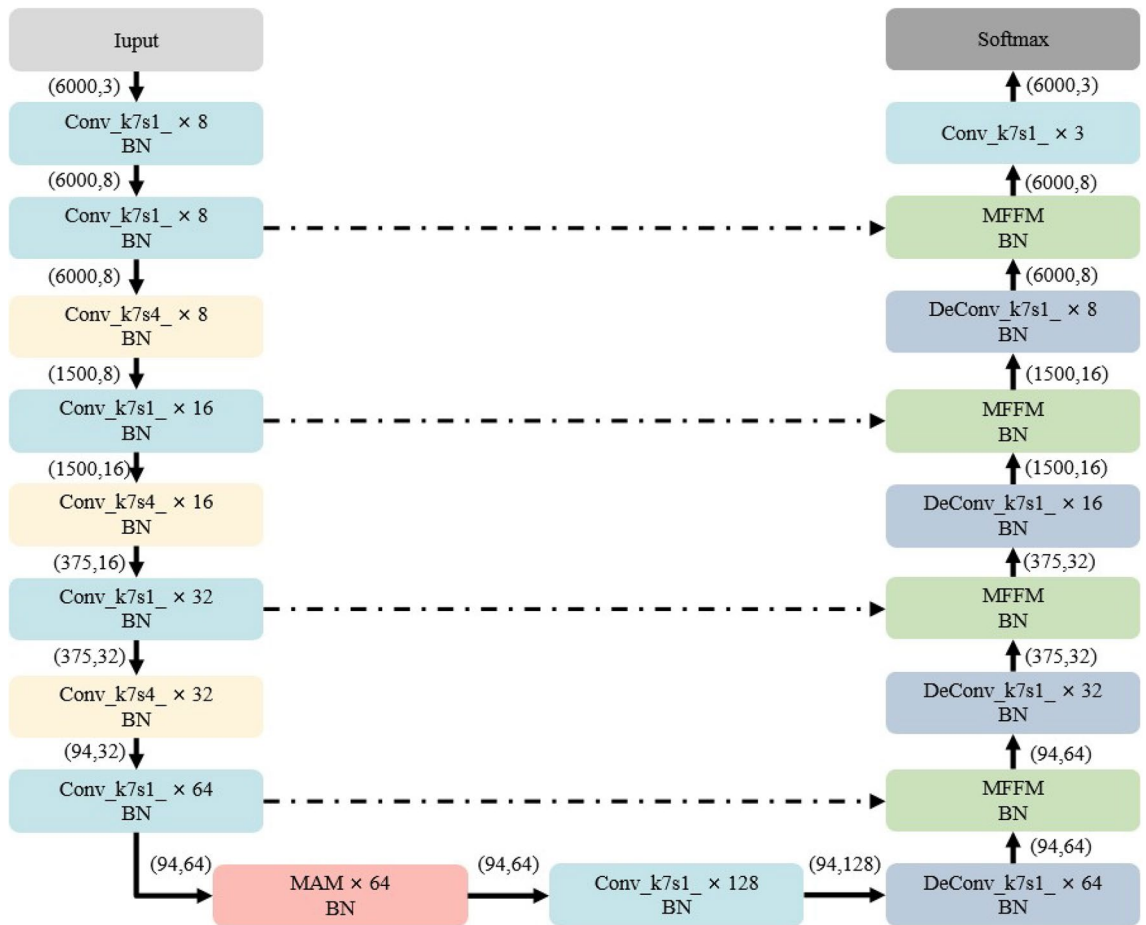
The remainder of this paper is organized as follows. “Proposed method” introduces the architecture of the proposed MFU-Net. “Experimental results and analysis” compares the accuracy of different seismic phase picking methods and presents ablation studies to evaluate the contribution of each key component in the proposed MFU-Net. “Conclusion” provides a conclusion of the work presented in this paper.

## Proposed method

In this section, we provide a detailed description of the key components of the proposed MFU-Net, including the overall architecture, the multi-scale feature fusion module, the multi-head attention mechanism<sup>22</sup>, and the loss function.

### Overall architecture

The overall architecture of the proposed MFU-Net is shown in Fig. 1. The output size is shown around each module. In the left part of Fig. 1, those 1D Convolution (Conv) layers are used to encode input signals and obtain multi-scale features. In the bottom part of Fig. 1, the MAM<sup>22</sup> and a Conv layer are the bottleneck layer of the proposed MFU-Net, which are used to perceive key signal regions. In the right part of Fig. 1, those Conv layers and DeConv layers are used to decode multi-scale features, while those MFFM<sup>21</sup> are used to fuse multi-scale



**Fig. 1.** Overall architecture of the proposed MFU-Net.  $k$  denotes the kernel size,  $s$  represents the stride, BN is the batch normalization,  $(i, j)$  indicates the signal length and the number of channels, Conv denotes a 1D standard convolution layer, and DeConv refers to transposed convolution. The MAM is the multi-head attention mechanism<sup>22</sup>, while the MFFM<sup>21</sup> denotes the multi-scale feature fusion module..

spatial and semantic features. Finally, the P-wave and S-wave arrival times are predicted by a Softmax operation. The MAM<sup>22</sup> and MFFM<sup>21</sup> will be detailed in the subsequent subsection.

**Multi-head attention mechanism**

The Multi-head Attention Mechanism<sup>22</sup> (MAM) is one of the core components of the transformer model. As shown in Fig. 2, it employs parallel “attention” operations (i.e., multi-head attention) and subsequently integrates their outputs, enabling the model to learn more comprehensive feature representations from diverse subspaces. Specifically, the MAM<sup>22</sup> first computes the query (Q), key (K), and value (V) through separate linear transformations. These linear projections map the original inputs into appropriate dimensions for attention calculation. Each head uses independent transformation parameters, allowing different heads to focus on different aspects of the input features. Formally, it can be expressed as:

$$Q_h = XW_h^Q, \quad K_h = XW_h^K, \quad V_h = XW_h^V, \tag{1}$$

where  $X$  is the input feature sequence.  $W_h^Q$ ,  $W_h^K$ , and  $W_h^V$  are the projection matrices for the  $h$ -th head. For each linearly transformed group  $Q_i$ ,  $K_i$ , and  $V_i$ , we perform the scaled dot-product attention operation as follows: First, compute the dot product (MatMul) between  $Q_i$  and  $K_i$  to obtain raw similarity scores. These scores are then scaled by a factor of  $1/\sqrt{d_k}$  (where  $d_k$  denotes the dimension of  $K_i$ ) to prevent gradient vanishing issues in the subsequent SoftMax caused by excessively large values. An optional masking operation may be applied before normalizing the scores through a SoftMax function to generate attention weights. Finally, the attention weights are multiplied with  $V_i$  (MatMul) to produce the output for the  $i$ -th attention head. The scaled dot-product attention can be formally expressed as:

$$\text{Attention}(Q_i, K_i, V_i) = \text{SoftMax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}} + \text{Mask}\right) V_i. \tag{2}$$

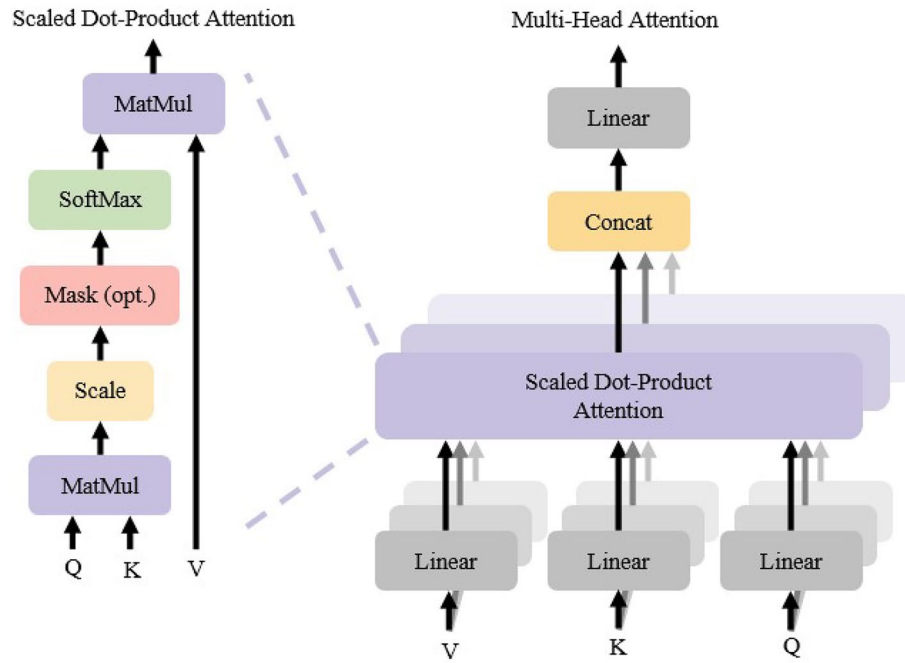


Fig. 2. Details of the multi-head attention mechanism<sup>22</sup>.

Finally, the outputs from all attention heads are concatenated and passed through a linear transformation to project the combined features into the desired output dimension, resulting in the final output of the multi-head attention mechanism<sup>22</sup>.

**Multi-scale feature fusion module**

The core idea of MFFM<sup>21</sup> is to employ convolution operations with different kernel sizes to capture features at multiple scales, where different kernel sizes represent different receptive fields. Specifically, there are three parallel branches consisting of 1D convolution with kernels sized 3×1, 5×1, and 7×1. This design is motivated by the multi-scale nature of seismic waveforms: smaller kernels (e.g., 3×1) are sensitive to sharp, localized features such as P-wave onsets, while larger kernels (e.g., 7×1) capture broader temporal structures like S-wave envelopes or coda waves. The intermediate kernel (5×1) provides complementary context between these extremes. This three-branch configuration balances representational capacity and model complexity, enabling effective multi-scale feature learning without excessive computational overhead.

By fusing these multi-dimensional features, the model can simultaneously leverage both fine-grained and coarse-grained information, thereby enhancing its ability to represent complex scenes and diverse targets, as shown in Fig. 3. Specifically, the process begins by performing a concatenation operation on two input features, merging the two inputs along the channel dimension. The formulation is as follows:

$$X_{\text{concat}} = \text{concat}(X_1, X_2), \tag{3}$$

where concat denotes the concatenation function applied along the channel dimension. Subsequently, three convolutional branches with different receptive fields are applied to extract the corresponding multi-scale features. Finally, the features extracted by the three branches ( $X_{\text{branch1}}$ ,  $X_{\text{branch2}}$ , and  $X_{\text{branch3}}$ ) are fused via element-wise addition to produce the final multi-scale feature representation. When fusion is performed through element-wise addition, the formulation is:

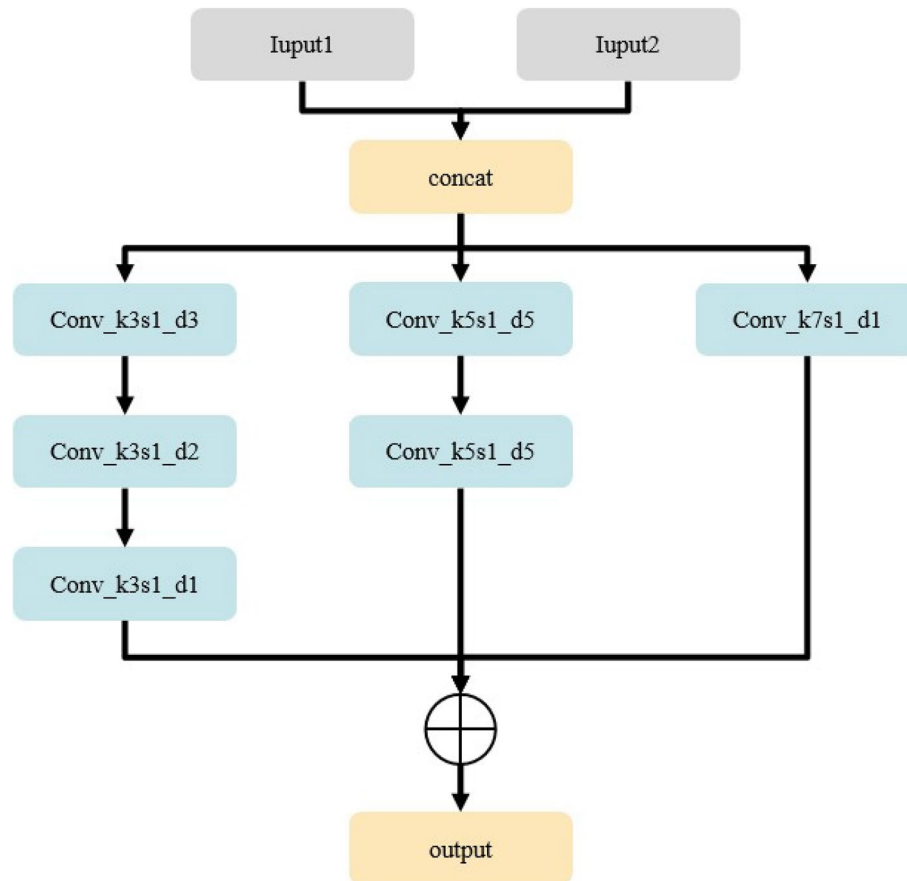
$$\text{output} = X_{\text{branch1}} + X_{\text{branch2}} + X_{\text{branch3}}. \tag{4}$$

**Loss function**

To address the severe class imbalance in the seismic phase picking task, we introduce a class-weighted loss function to mitigate bias toward majority classes, formulated as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c \in \mathcal{C}} w_c \cdot y_{i,c} \log(p_{i,c}) \tag{5}$$

where  $\mathcal{C} = \{\text{P-wave, S-wave, noise}\}$  denotes the class set.  $y_{i,c}$  and  $p_{i,c}$  are the ground truth and predicted probability for the class  $c$  at the sample  $i$ . The  $w_c$  is the class-specific weight inversely proportional to class frequency. Specifically, let  $f_c$  denote the frequency of class  $c$  in the training set, it is defined as:



**Fig. 3.** Details of the multi-scale feature fusion module<sup>21</sup>.

$$f_c = \frac{n_c}{N}, \quad (6)$$

where  $n_c$  is the number of samples belonging to class  $c$ . The class weight  $w_c$  is then defined as inversely proportional to  $f_c$ :

$$w_c = \frac{1}{f_c}. \quad (7)$$

To stabilize training and maintain the overall scale of the loss, we normalize the weights such that their mean equals 1 (or equivalently, the sum equals the number of classes  $C$ ):

$$w_c = \frac{1/f_c}{\frac{1}{C} \sum_{k=1}^C (1/f_k)}. \quad (8)$$

This formulation ensures that when all classes are balanced ( $f_c = 1/C$  for all  $c$ ), all weights equal 1.

## Experimental results and analysis

To demonstrate the superior performance of our method, we conduct qualitative evaluation, quantitative assessment, and ablation study, respectively.

### Experiment settings

Our model is implemented on the PyTorch 2.1.1 framework with an Inter(R) i7-12700K CPU, 64GB RAM, and an NVIDIA RTX 4080 GPU. The ADAM optimizer is used to optimize the model's weights, and the initial learning rate is set to 0.001. The batch size and training epochs are set to 64 and 300, respectively.

#### Comparison between datasets

The Fujian dataset is characterized by its regional focus, lower event count, and higher peak amplitudes, making it suitable for evaluating performance on local, high-amplitude events. In contrast, the STEAD dataset offers global diversity, a much larger sample size, and a broader magnitude range, but with lower overall amplitude

values due to normalization or different recording scales. The combination of these two datasets allows our model to learn robust features applicable to both local and global seismic scenarios.

## Datasets

We utilize two datasets from the Fujian Earthquake Agency and the STEAD dataset<sup>23</sup> in this study. During the experiments, two datasets are divided into training, validation, and test sets in a ratio of 8:1:1.

- (1) *Dataset from the Fujian Earthquake Agency*: this dataset consists of strong-motion acceleration waveform records collected over the past decade, provided by the Fujian Earthquake Agency. It contains a total of 1,059 three-component waveform samples (Z, N, E components), with a sampling rate of 100 Hz. After splitting according to the 8:1:1 ratio, the final dataset comprises: Training set: 1,754 samples, Validation set: 219 samples, and Test set: 220 samples. The maximum absolute amplitude values across all waveforms range from 0.165 to 577,557, with a mean value of approximately 2,676.
- (2) *STEAD Dataset*: developed by Stanford University, the STEAD dataset comprises high-quality global seismic and non-seismic signals, recorded at a sampling rate of 100 Hz across three orthogonal components. Each waveform includes precise annotations for P-wave and S-wave arrival times, spanning a wide range of regions and magnitudes, making it highly diverse. To enhance the clarity and quality of seismic features in the training data, only events with an epicentral distance within 100 km are selected for this study. After applying the 8:1:1 split to the filtered subset, the final dataset comprises: Training set: 101,2525 samples, Validation set: 12,6565 samples, and Test set: 12,6567 samples. The maximum absolute amplitude values range from 1.65 to 32.25, with a mean value of approximately 9.04.

## Evaluation metrics

To comprehensively evaluate all models' performance, we adopted three evaluation metrics, including Precision, Recall, and F1-score. Precision measures the proportion of correctly predicted positive samples among all samples predicted as positive by the model. Recall focuses on the proportion of actual positive samples that are correctly identified by the model. F1-score is the harmonic mean of precision and recall, providing a balanced metric that takes both into account. It is particularly useful for evaluating the model's performance in classification tasks where a trade-off between precision and recall is important. The formula of these three metrics is shown as follows:

$$\begin{aligned} \text{Precision} &= TP / (TP + FP), \\ \text{Recall} &= TP / (TP + FN), \\ \text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \end{aligned} \quad (9)$$

where true positives  $TP$  refer to the number of positive samples correctly identified by the model, false positives  $FP$  are negative samples incorrectly classified as positive, and false negatives  $FN$  represent positive samples that the model failed to detect. In this task, a tolerance of 0.1 seconds is applied to account for small timing discrepancies. Under this setting,  $FP$  is defined as samples whose residuals fall within this tolerance window but are incorrectly predicted. As the harmonic mean of precision and recall,  $F1$  provides a balanced view of the model's performance, especially under conditions where both false positives and false negatives carry significant importance.

## Comparison of different picking methods

To comprehensively evaluate the performance of the phase picking model, we conducted comparisons with several state-of-the-art approaches, including SEANet<sup>14</sup>, SegPhase<sup>24</sup>, and GPD<sup>25</sup>. All models were trained and evaluated on the same datasets to ensure fairness and comparability across experiments. During training, data augmentation techniques such as random cropping and time shifting were employed to accelerate convergence, reduce overfitting, and enhance the model's generalization capability. The test data remained unfiltered to preserve its authenticity and reflect real-world performance.

The quantitative result is shown in Tables 1 and 2. As shown in Table 1, our model outperforms SEANet, GPD, and SegPhase in all evaluation indicators. For the P-wave, the Precision of our model reaches 0.972, which is 0.018, 0.015, and 0.028 higher than that of SEANet (0.954), GPD (0.957), and SegPhase (0.944), respectively. The Recall of 0.959 is also better than the other three models, and the F1-score of 0.965 is significantly ahead, indicating that our model has a more accurate and comprehensive identification ability for P-wave. In terms of the S-wave, the advantages of our model are more prominent. Its Precision is 0.984, far exceeding SEANet

Model	P-wave			S-wave		
	Precision	Recall	F1	Precision	Recall	F1
SEANet	0.954	0.958	0.956	0.931	0.629	0.751
GPD	0.957	0.945	0.958	0.945	0.846	0.893
SegPhase	0.944	0.949	0.948	0.944	0.827	0.882
Ours	0.972	0.959	0.965	0.984	0.888	0.934

**Table 1.** Quantitative results on the dataset from the Fujian Earthquake Agency.

Model	P-wave			S-wave		
	Precision	Recall	F1	Precision	Recall	F1
SEAnet	0.958	0.956	0.957	0.949	0.729	0.824
GPD	0.963	0.956	0.959	0.953	0.845	0.896
SegPhase	0.957	0.955	0.956	0.955	0.895	0.924
Ours	0.976	0.956	0.965	0.981	0.900	0.941

**Table 2.** Quantitative results on the STEAD dataset from Stanford University.

(0.931), GPD (0.945), and SegPhase (0.944). The Recall of 0.888 and the F1-score of 0.934 are also the highest among all models, which means that our model can effectively identify S-wave while reducing misjudgment and missed judgment, showing strong robustness and superiority in the task of seismic wave identification.

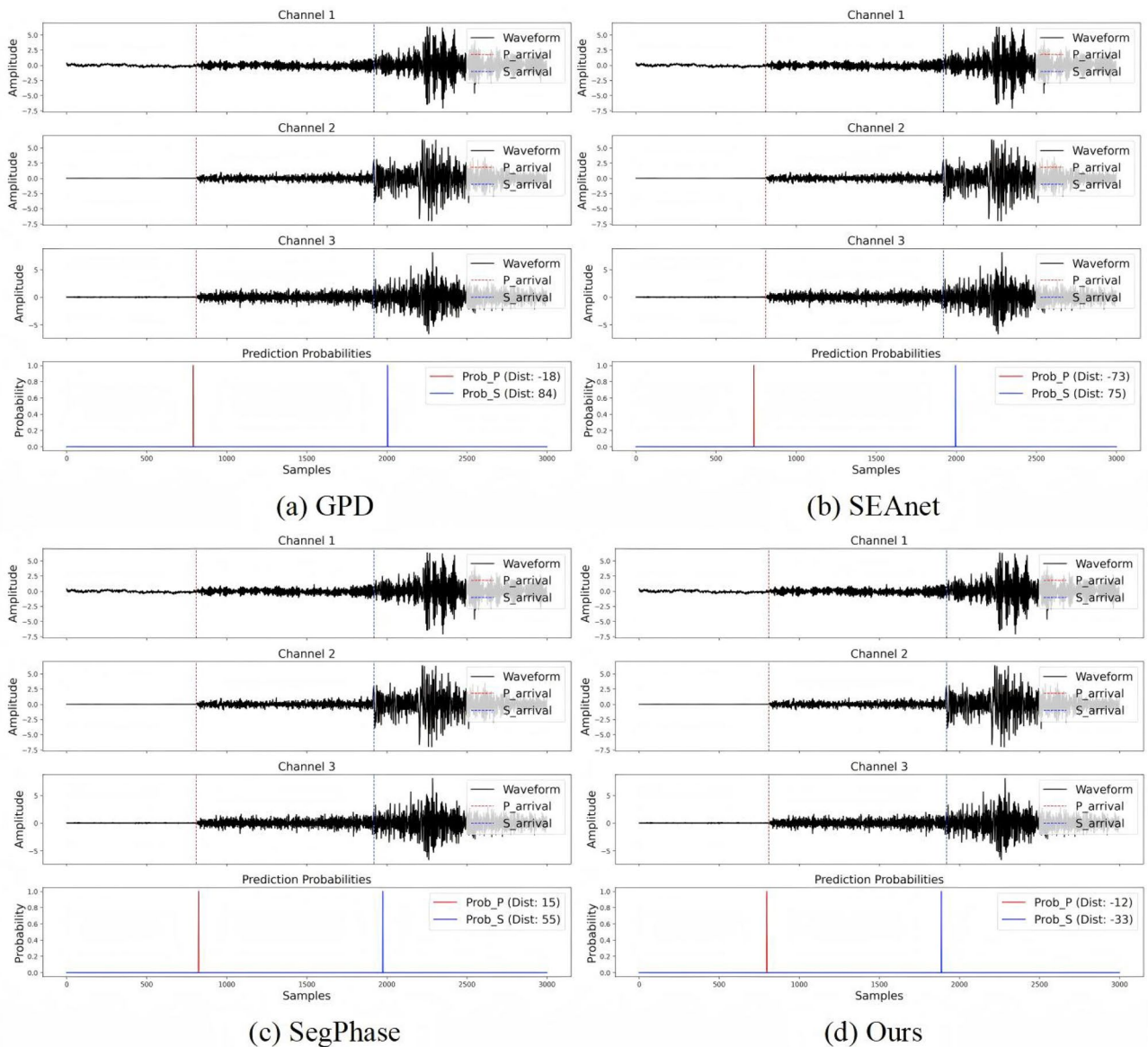
Table 2 reports the quantitative results of multiple models on the STEAD dataset from Stanford University, evaluating their performance in identifying P-wave and S-wave. Among the compared models (SEAnet, GPD, SegPhase), our model demonstrates significant advantages across all metrics. For the P-wave, the Precision of our model reaches 0.976, outperforming SEAnet (0.958), GPD (0.963), and SegPhase (0.957). While sharing the same Recall of 0.956 as GPD and SEAnet, our model achieves a higher F1-score of 0.965. For the S-wave, our model shows even more remarkable performance: its Precision of 0.981, Recall of 0.900, and F1-score of 0.941 are all the highest among all models. Specifically, compared with SegPhase (the second-best in S-wave), our F1-score is 0.017 higher. These results fully illustrate that our model has stronger adaptability and higher recognition accuracy on the STEAD dataset, effectively improving the performance of seismic wave identification.

Further, we provide partial visualizations of the prediction results, as shown in Figs. 4 and 5. As shown Fig. 4, the GPD model<sup>25</sup> can capture the general trend of seismic waves, but, the prediction probability curve shows certain fluctuations. In its channel 2, the identification of the S-wave arrival time has a deviation, which affects the overall accuracy of seismic phase recognition. As shown in Figs. 5b, the SEAnet model<sup>14</sup> are relatively stable for the waveform identification results on each channel, but in Channel 3, its prediction probability for the P-wave has a slow rise, leading to a delay in the determined arrival time to some extent, and the accuracy of S-wave identification needs to be improved. The SegPhase model<sup>24</sup> are obvious misalignments in the identification of P-wave and S-wave arrival times in partial channels (e.g., channel 2), as depicted in Fig. 5. The prediction probability curve is not smooth enough, indicating that the model has limitations in capturing the accurate characteristics of seismic waves. In contrast, our model shows significant advantages. On all three channels, the prediction probability curves rise rapidly and stably at the actual arrival times of P-wave and S-wave, and the determined arrival times (red/blue dashed lines) are highly consistent with the true positions of seismic phases. Whether in the capture of weak seismic signals (channel 1) or the identification of complex waveform segments (channel 3), our model can accurately locate the arrival times of P-wave and S-waves, and the prediction probability curves are smooth and reliable. These visual results fully demonstrate that our model has higher accuracy and robustness in seismic wave phase identification. Compared with GPD<sup>25</sup>, SEAnet<sup>14</sup>, and SegPhase<sup>24</sup>, it can more effectively capture the characteristic information of seismic waves and provide more accurate seismic phase arrival time identification, which is of great significance for improving the performance of seismic event analysis.

To provide a more comprehensive evaluation of the deep learning models, we have expanded our analysis by including representative examples across different noise levels and event magnitudes. As shown in Fig. 6, under low-noise conditions, all models perform reliably, serving as a baseline for comparison. For high-noise scenarios, Fig. 7 reveals that our model maintains robust performance, outperforming GPD and SEAnet, which exhibit significant mis-picks. In contrast, Fig. 8 illustrates the models' performance on a low-magnitude event, where our proposed model demonstrates superior sensitivity by accurately detecting both P- and S-phases, while SegPhase fails to identify the S-phase entirely. Finally, Fig. 9 shows that even for high-magnitude events, our model provides the most precise picks with minimal timing errors. These additional comparisons provide a holistic view of each model's strengths and weaknesses under diverse real-world conditions.

In addition, Fig. 10 illustrates the prediction results of our method along with the corresponding attention regions, shedding light on the model's feature-capturing mechanism in seismic wave recognition. In Fig. 10a, b, our method accurately identifies the P-wave and S-wave arrival times across all three channels. The predicted probability curves (e.g., red for P-waves and blue for S-waves) exhibit sharp rises at the true arrival points, with clearly defined boundaries—indicating high confidence and precise detection. Our model's attention (highlighted by colored regions) is precisely focused on the segments containing P- and S-waves. For the P-wave (pink region), attention is tightly concentrated around the actual arrival time, effectively capturing the transient and subtle signal without being disturbed by background noise. For the S-wave (blue region), attention spans the entire oscillation period, ensuring comprehensive feature extraction from the longer and more complex waveform. These results demonstrate that our method possesses an inherent mechanism for adaptive feature focusing, which contributes to its enhanced robustness and accuracy in seismic phase identification.

To provide a more intuitive understanding of the performance differences among the compared methods, we present a comparison of their mean absolute errors (MAE) for P-wave and S-wave picking in Fig. 11. As shown in Fig. 11, our proposed model (PhaseNet) achieves the lowest MAE for both P-wave (13.3 samples) and S-wave (36.1 samples) picking, significantly outperforming GPD, SEAnet, and SegPhase. The large MAE values for other methods (e.g., GPD with 1122.1 and 597.1 samples for P- and S-waves, respectively) indicate that they suffer from frequent large picking errors. In contrast, PhaseNet's consistently low MAE suggests its predictions



**Figure 4.** Visualization of prediction on the dataset from the Fujian Earthquake Agency..

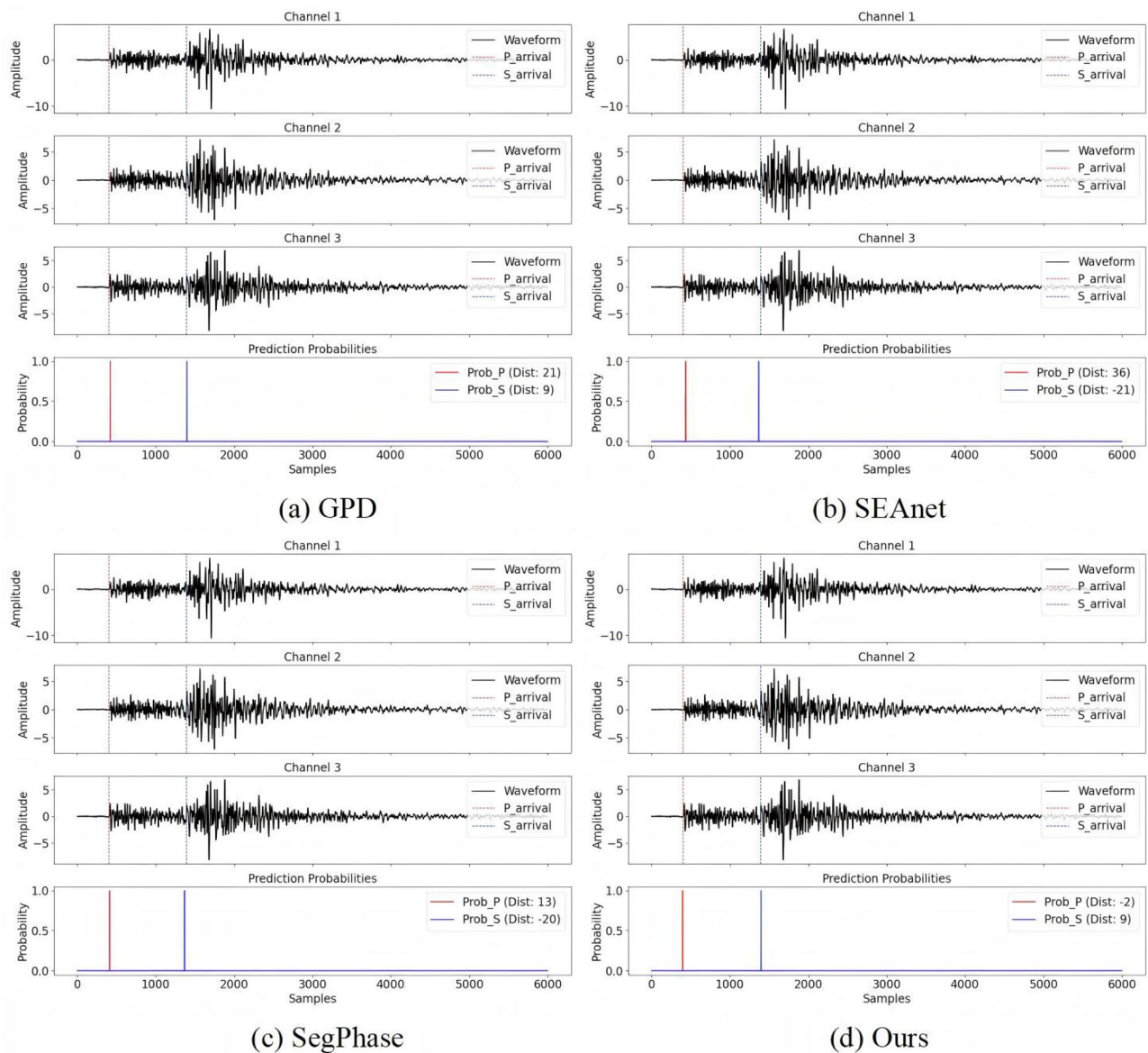
are not only accurate on average but also exhibit high stability with minimal outliers. This robustness is critical for reliable real-time seismic monitoring applications.

**Ablation study**

To demonstrate the effect of each component in our method, we conduct an ablation study, including the following experiments: (1) our method without MAM module (w/o MAM), (2) our method without MFFM module (w/o MFFM), (3) our method without the class-weighted loss (w/o CWL), and (4) our full model (Full model). The quantitative comparison on tow datasets is shown in Tables 3 and 4.

Table 3 presents the quantitative results of the ablation study on the dataset from the Fujian Earthquake Agency, aiming to verify the contribution of key components (MAM, MFFM, CWL) to our model’s performance in P-wave and S-wave identification. For the P-wave, w/o CWL even fails completely (all metrics are 0.000), indicating the critical role of CWL in P-wave identification. w/o MAM and w/o MFFM models have Precision values of 0.804 and 0.842, Recall values of 0.916 and 0.941, and F -scores of 0.857 and 0.888, respectively. In contrast, full model achieves a Precision of 0.900, a Recall of 0.947, and an F1-score of 0.923, significantly outperforming the models missing individual components. For the S-wave, w/o CWL model also fails (all metrics are 0.000), while the w/o MAM and w/o MFFM models have relatively low performance (e.g., w/o MFFM has a Precision of 0.753 and an F1-score of 0.794). The full model achieves a Precision of 0.864, a Recall of 0.853, and an F1-score of 0.858, demonstrating a more stable and excellent identification ability.

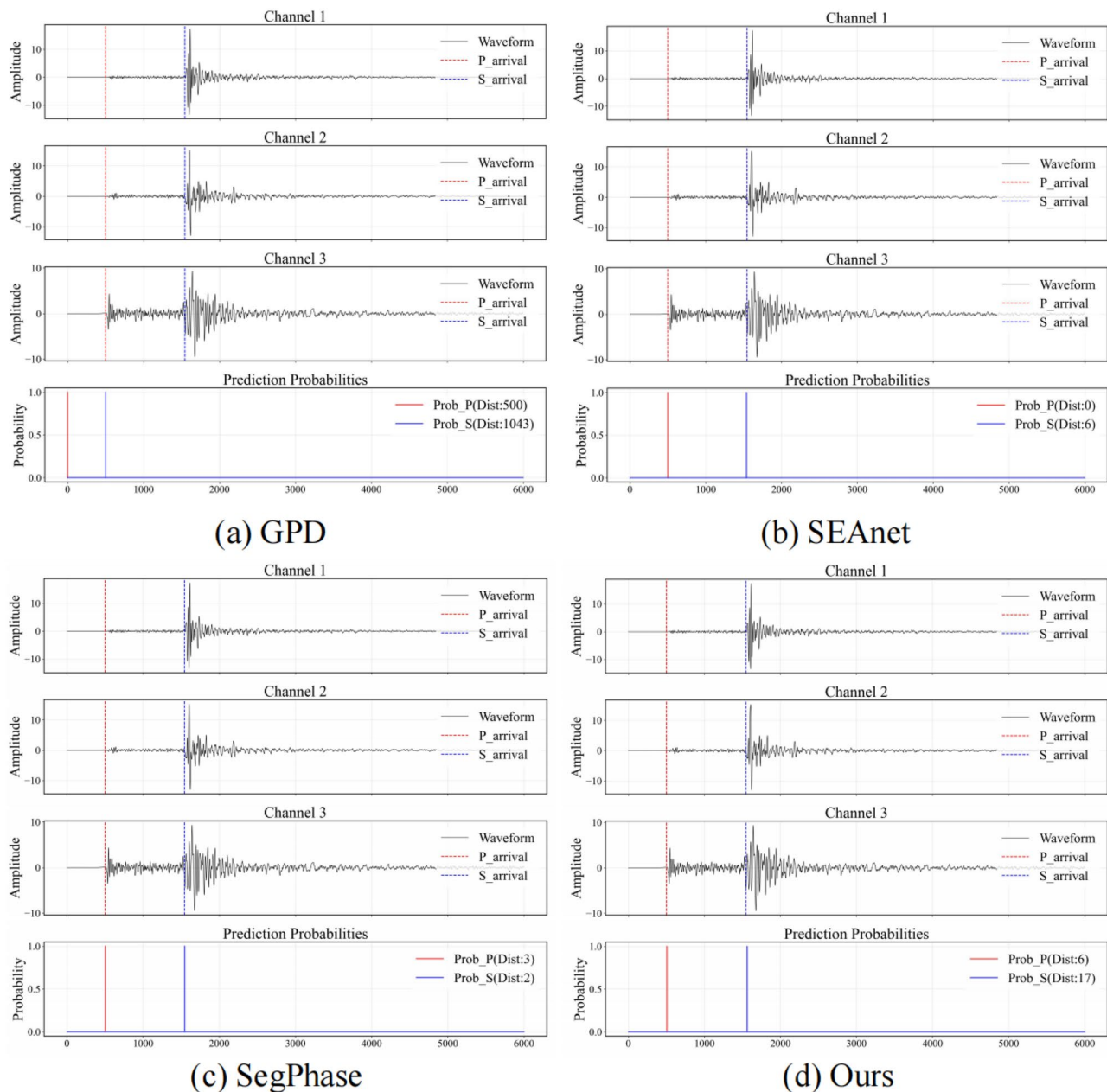
As shown in Table 4, the full model significantly outperforms variants missing individual components (w/o MAM, w/o MFFM, w/o CWL) across all metrics. For the P-wave, the w/o CWL model completely fails (all



**Fig. 5.** Visualization of prediction on the STEAD dataset from Stanford University.

metrics = 0.000), highlighting CWLs critical role. w/o MAM and w/o MFFM models achieve Precision values of 0.969 and 0.950, Recall values of 0.957 and 0.956, and F1-scores of 0.963 and 0.953, respectively. In contrast, the full model reaches a Precision of 0.972, a Recall of 0.959, and an F1-score of 0.965, which demonstrates superior accuracy and robustness. For the S-wave, w/o CWL model also fails (all metrics = 0.000). While w/o MAM and w/o MFFM show partial effectiveness (e.g., w/o MFFM has an F1-score of 0.920), the full model achieves a Precision of 0.984, a Recall of 0.888, and an F1-score of 0.934.

The reported zero metrics in the w/o CWL setting are not due to a complete failure of the model to produce any meaningful output, but rather stem from our specific evaluation protocol. In our implementation, a detection is counted as a true positive only if: (1) the predicted probability at the peak location of the P- or S-wave channel is the global maximum across the entire time series for that channel, and (2) this peak probability is strictly greater than the corresponding predictions in all other channels at the same time step (i.e., the model must exhibit clear class-wise discriminability). When CWL is removed, the model's predictions for seismic phases become diffuse and lack sufficient confidence or inter-class separation. Consequently, no prediction satisfies the above criteria, and all detections are treated as negatives—leading to zero precision, recall, and F1 scores under this strict evaluation scheme. These results confirm that each component (MAM, MFFM, CWL) contributes uniquely, and their integration in the full model yields optimal performance. The ablation study validates that our complete method excels on the dataset from the Fujian Earthquake Agency and the STEAD dataset, showcasing stronger adaptability and higher identification accuracy for seismic waves.



**Fig. 6.** Visualization of prediction on the low noise signal from the STEAD dataset..

### Computational efficiency

To evaluate the practicality of our proposed model for real-time seismic phase picking, we conduct a comprehensive analysis of its computational cost and inference speed. We compare our model against three representative state-of-the-art methods: GPD, SEANet, and SegPhase. The comparison metrics include the number of model parameters (in millions, M), floating-point operations (FLOPs, in Giga), and average inference time per waveform (in seconds, s).

The results are summarized in Table 5. Our model, despite achieving high performance with minimal architectural modifications, has a parameter count of 0.837M, which is higher than SEANet (0.074M) but comparable to or lower than GPD (0.512M) and SegPhase (0.331M). In terms of computational complexity, our model requires only 0.242G FLOPs, significantly less than both GPD (0.499G) and SegPhase (0.533G), and slightly more than SEANet (0.020G). Crucially, our model demonstrates competitive inference speed, with an average latency of 0.008s per waveform, which is faster than SEANet (0.015s) and SegPhase (0.009s), and only marginally slower than GPD (0.002s).

These results indicate that while our model is not the lightest in terms of parameters, it achieves an excellent balance between accuracy and efficiency. Its low FLOP count and fast inference speed make it highly suitable for real-time applications where rapid processing is critical.

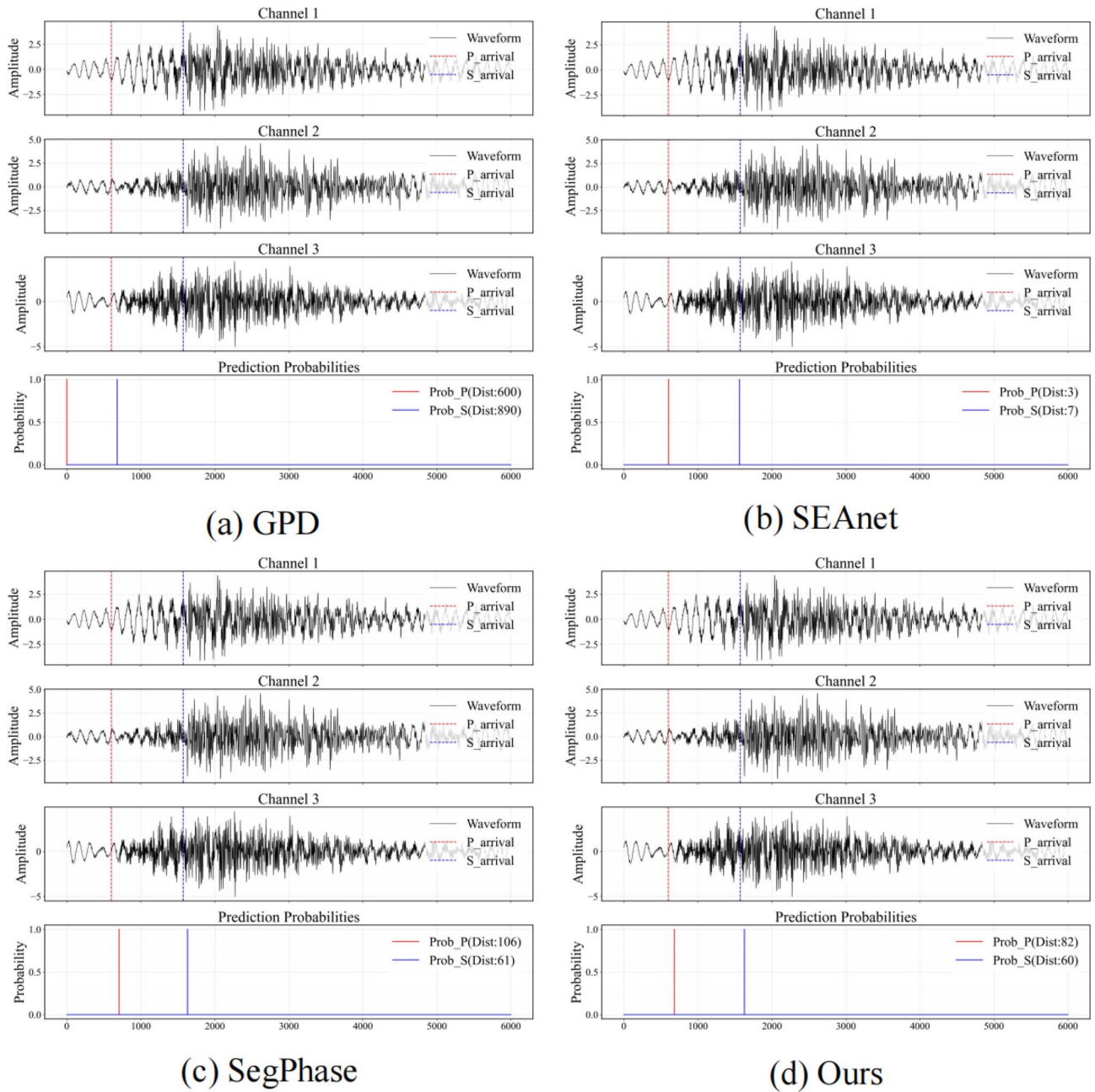


Fig. 7. Visualization of prediction on the high noise signal from the STEAD dataset..

### Conclusion

This study proposes a seismic phase picking network that enhances the feature fusion and bottleneck components of the U-Net architecture, aiming to achieve optimal prediction performance with minimal architectural modifications. In addition, a class-balanced loss function is introduced to effectively guide the model’s attention toward underrepresented waveform classes. Both qualitative and quantitative experimental comparisons with several state-of-the-art methods demonstrate that the proposed MFU-Net achieves superior performance. The findings of this work offer a novel and effective direction for developing efficient seismic phase picking networks.

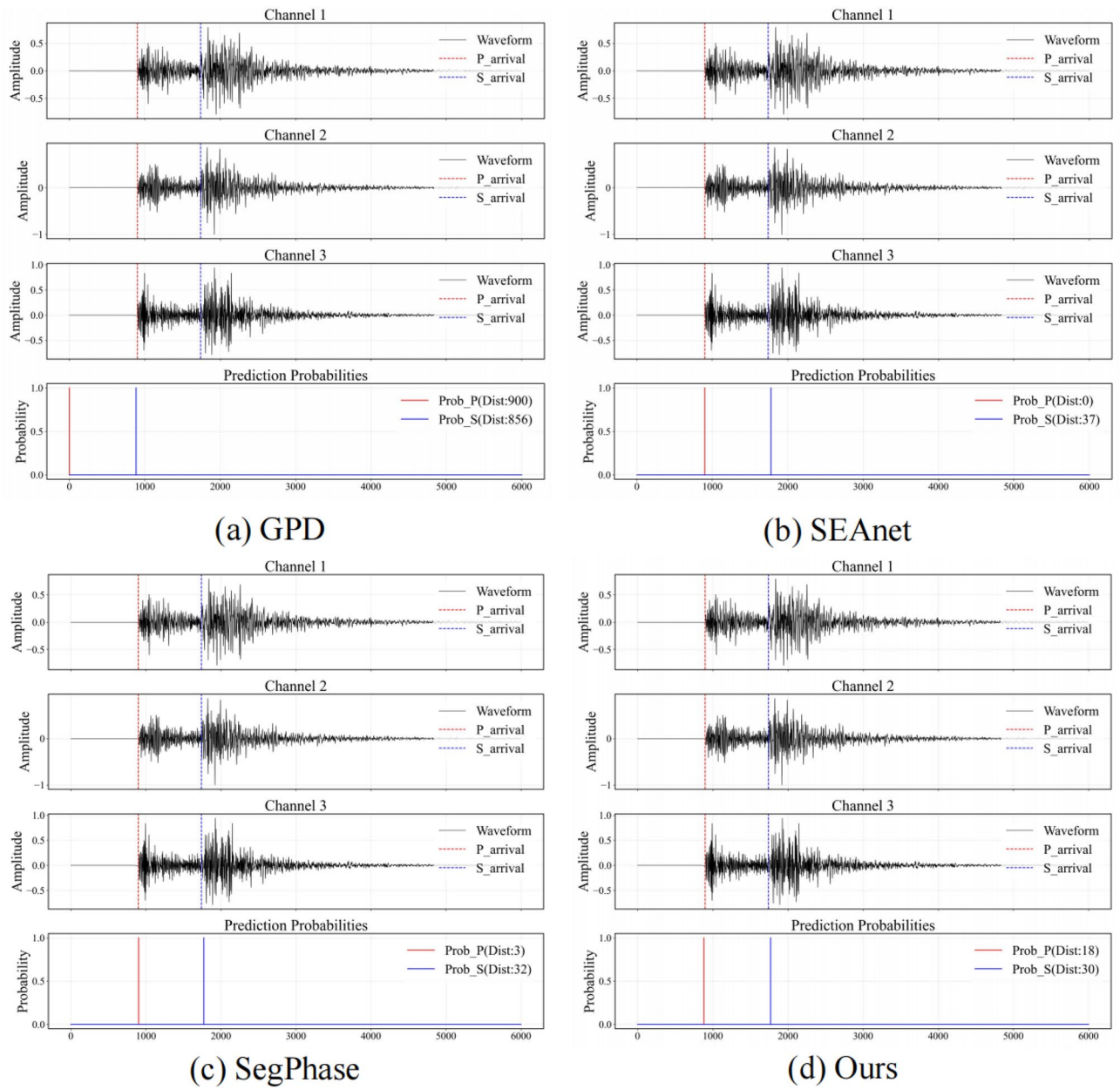
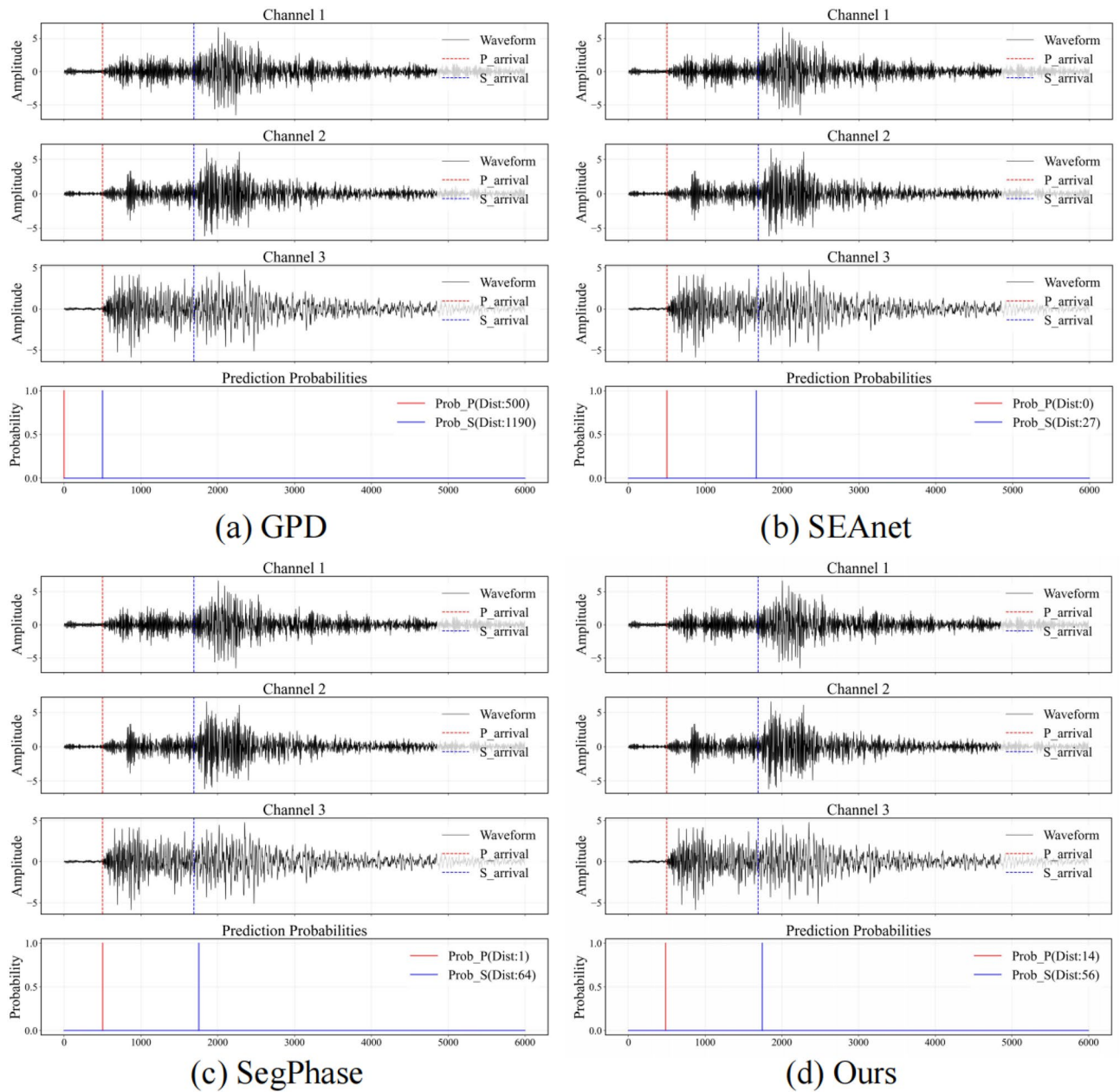
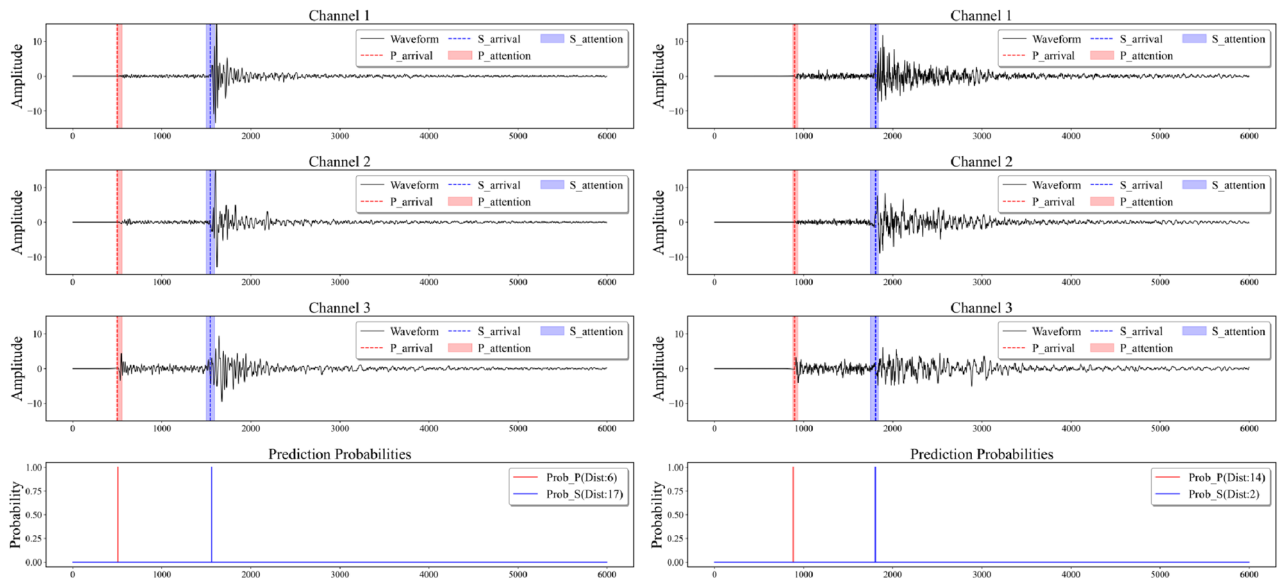


Fig. 8. Visualization of prediction on the low-amplitude signal from the STEAD dataset..



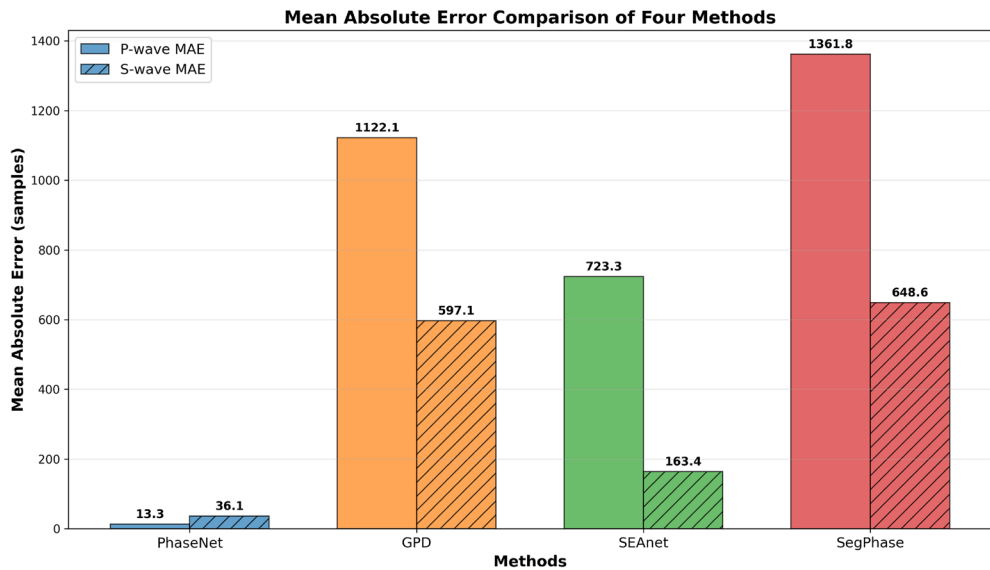
**Fig. 9.** Visualization of prediction on the high-amplitude signal from the STEAD dataset..



(a) Predicted Result 1

(b) Predicted Result 2

**Fig. 10.** Visualization of the attention area for our model. Corresponding Attention Area, highlighting the regions where the model focuses on extracting features.



**Fig. 11.** Mean absolute error comparison for P-wave and S-wave picking across four different methods. Lower values indicate better performance..

Model	P-wave			S-wave		
	Precision	Recall	F1	Precision	Recall	F1
w/o MAM	0.804	0.916	0.857	0.855	0.820	0.837
w/o MFFM	0.842	0.941	0.888	0.753	0.840	0.794
w/o CWL	0.000	0.000	0.000	0.000	0.000	0.000
Full model	0.900	0.947	0.923	0.864	0.853	0.858

**Table 3.** Quantitative results of the ablation study on the dataset from the Fujian Earthquake Agency.

Model	P-wave			S-wave		
	Precision	Recall	F1	Precision	Recall	F1
w/o MAM	0.969	0.957	0.963	0.977	0.816	0.889
w/o MFFM	0.950	0.956	0.953	0.974	0.871	0.920
w/o CWL	0.000	0.000	0.000	0.000	0.000	0.000
Full model	0.972	0.959	0.965	<b>0.984</b>	0.888	0.934

**Table 4.** Quantitative results of the ablation study on the STEAD dataset from Stanford University.

Model	Parameters (M)	FLOPs (G)	Inference Time (s)
GPD	0.512	0.499	0.002
SEANet	0.074	0.020	0.015
SegPhase	0.331	0.533	0.009
Ours	0.837	0.242	0.008

**Table 5.** Comparison of computational efficiency among different models.

## Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 15 September 2025; Accepted: 8 December 2025

Published online: 17 December 2025

## References

- Zhu, D., Li, Y. & Zhang, C. Automatic time picking for microseismic data based on a fuzzy c-means clustering algorithm. *IEEE Geosci. Remote Sens. Lett.* **13**, 1900–1904 (2016).
- Luo, H. et al. The CGAS deep learning algorithm for p-wave arrival time picking of mining microseismic events. *IEEE Access* **11**, 102961–102970 (2023).
- Qin, Z., Pan, S., Chen, J., Cui, Q. & He, J. Method of automatically detecting the abnormal first arrivals using delay time (December 2020). *IEEE Trans. Geosci. Remote Sens.* **60**, 1–8 (2022).
- Allen, R. Automatic phase pickers: Their present use and future prospects. *Bull. Seismol. Soc. Am.* **72**, 225–242 (1982).
- Sleeman, R. & Van Eck, T. Robust automatic p-phase picking: An on-line implementation in the analysis of broadband seismogram recordings. *Phys. Earth Planet. Interiors* **113**, 265–275 (1999).
- Yang, H., Zhu, L. & Chu, R. Fault-plane determination of the 18 April 2008 Mount Carmel, Illinois, earthquake by detecting and relocating aftershocks. *Bull. Seismol. Soc. Am.* **99**, 3413–3420 (2009).
- Zhu, W. & Beroza, G. C. PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophys. J. Int.* **216**, 261–273 (2019).
- Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *Proc. Med Image Comput Assist Interv.* 234–241 (2015).
- Saad, O. M., Inoue, K., Shalaby, A., Samy, L. & Sayed, M. S. Automatic arrival time detection for earthquakes based on stacked denoising autoencoder. *IEEE Geosci. Remote Sens. Lett.* **15**, 1687–1691 (2018).
- Li, B., Fan, L., Jiang, C., Liao, S. & Fang, L. CSESnet: A deep learning p-wave detection model based on U-Net++ designed for China seismic experimental site. *Front. Earth Sci.* **10-2022** (2023).
- Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **39**, 1856–1867 (2019).
- Ma, Y., Cao, S., Rector, J. W. & Zhang, Z. Automated arrival-time picking using a pixel-level network. *Geophysics* **85**, 415–423 (2020).
- Wang, J., Xiao, Z., Liu, C., Zhao, D. & Yao, Z. Deep learning for picking seismic arrival times. *J. Geophys. Res. Solid Earth* **124**, 6612–6624 (2019).
- Hou, X., Zheng, Y., Jiang, M. & Zhang, S. SEA-net: Sequence attention network for seismic event detection and phase arrival picking. *Eng. Appl. Artif. Intell.* **122**, 106090 (2023).
- Li, W. et al. EPick: Attention-based multi-scale U-Net for earthquake detection and seismic phase picking. *Front. Earth Sci.* **10**, 953007 (2022).
- Chai, C. et al. Using a deep neural network and transfer learning to bridge scales for seismic phase picking. *Geophys. Res. Lett.* **47**, e2020GL088651 (2020).
- Lan, B. et al. Seismic phase picking using a cross-attention network on Nvidia Jetson Xavier NX. *IEEE Access* (2024).
- Zhu, Y., Zhao, S., Wei, W., Zhang, F. & Zhao, F. A lightweight network for seismic phase picking on embedded systems. *IEEE Access* **12**, 85103–85114 (2024).
- Yu, Z., Wang, W. & Chen, Y. Benchmark on the accuracy and efficiency of several neural network based phase pickers using datasets from china seismic network. *Earthq. Sci.* **36**, 113–131 (2023).
- Liao, S. et al. Development of a real-time intelligent seismic processing system and its application in the 2021 Yunnan Yangbi m<sub>s</sub>6.4 earthquake. *Chin. J. Geophys. (in Chinese)* **64**, 3632–3645 (2021).
- Cui, Y., Bai, M., Wu, J. & Chen, Y. Earthquake signal detection using a multiscale feature fusion network with hybrid attention mechanism. *Geophys. J. Int.* **240**, 988–1008 (2024).
- Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* (Guyon, I. et al. eds.). Vol. 30 (2017).
- Mousavi, S. M., Sheng, Y., Zhu, W. & Beroza, G. C. Stanford earthquake dataset (stead): A global data set of seismic signals for AI. *IEEE Access* **7**, 179464–179476 (2019).

24. Katoh, S. et al. SegPhase: Development of arrival time picking models for Japan's seismic network using the hierarchical vision transformer. *Earth Planets Sp.* **77**, 118 (2025).
25. Ross, Z. E., Meier, M., Hauksson, E. & Heaton, T. H. Generalized seismic phase detection with deep learning. *Bull. Seismol. Soc. Am.* **108**, 2894–2901 (2018).

### Acknowledgements

This work is supported by National Natural Science Foundation of China (No.52378523), Fujian Science and Technology Programme (No. 2025Y0047).

### Author contributions

Conceived and designed the experiments: L.W., J.Y., and S.L. Performed the experiments: L.W., Y.Z., M.H., and H.Z. Analyzed the data: J.Y. and S.L. Wrote and reviewed the paper: L.W., J.Y., Y.Z., M.H., H.Z., and S.L.

### Funding

This work belongs to the project Grant No. 52378523 funded by the National Natural Science Foundation, and the project Grant No. 2025Y0047 funded by the Fujian Science and Technology Programme.

### Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to S.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025