SeisT: A foundational deep learning model for earthquake monitoring tasks

Sen Li[1], Xu Yang[1*], Anye Cao[2*], Changbin Wang[3], Yaoqi Liu[2], Yapeng Liu[1], Qiang Niu[1]

[1*]School of Computer Science and Technology, China University of Mining and Technology, Daxue Road, Xuzhou, 221116, Jiangsu, China.
[2]School of Mines, China University of Mining and Technology, Daxue Road, Xuzhou, 221116, Jiangsu, China.
[3]State Key Laboratory of Coal Resources and Safe Mining, China University of Mining and Technology, Daxue Road, Xuzhou, 221116, Jiangsu, China.

*Corresponding author(s). E-mail(s): yang_xu@cumt.edu.cn; caoanye@163.com;
Contributing authors: sli@cumt.edu.cn; changbin.wang@cumt.edu.cn;
yaoqi_liu@cumt.edu.cn; liuyp@cumt.edu.cn; niuq@cumt.edu.cn;

Abstract

Seismograms, the fundamental seismic records, have revolutionized earthquake research and monitoring. Recent advancements in deep learning have further enhanced seismic signal processing, leading to even more precise and effective earthquake monitoring capabilities. This paper introduces a foundational deep learning model, the Seismogram Transformer (SeisT), designed for a variety of earthquake monitoring tasks. SeisT combines multiple modules tailored to different tasks and exhibits impressive out-of-distribution generalization performance, outperforming or matching state-of-the-art models in tasks like earthquake detection, seismic phase picking, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation. The performance scores on the tasks are 0.96, 0.96, 0.68, 0.95, 0.86, 0.55, and 0.81, respectively. The most significant improvements, in comparison to existing models, are observed in phase-P picking, phase-S picking, and magnitude estimation, with gains of 1.7%, 9.5%, and 8.0%, respectively. Our study, through rigorous experiments and evaluations, suggests that SeisT has the potential to contribute to the advancement of seismic signal processing and earthquake research.

Keywords: Seismogram, deep learning, phase picking, magnitude estimation, back-azimuth estimation, first-motion polarity classification, epicentral distance estimation

1 Introduction

Deep learning approaches have gained traction in various seismological subfields, exhibiting promising capabilities that often surpass traditional methods and lead to notable performance enhancements [1–5]. Their efficiency and robust performance have made deep learning-based tools in seismology increasingly popular in various research domains.

Recorded seismic waveforms serve as the starting point for seismicity monitoring and form the basis for a series of subsequent analyses. Deep learning models acquire compact representations of seismic signals from large-volume seismic datasets, enabling them to extract deep-level feature representations of seismic records. Notably, in the domain of seismology, significant advancements have been achieved through deep learning methodologies applied to single-station scenarios [6]. These advancements encompass diverse aspects, including earthquake detection [7–10], phase picking [8–12], first-motion polarity

classification [13–15], magnitude estimation [16, 17], back-azimuth estimation [18, 19], and epicentral distance estimation [18, 20]. However, research on foundational models in seismology is still in its early stages, with only a few relevant models recently proposed, such as SeisCLIP [21] and SFM [22]. A foundational model [23], trained on large-volume seismic datasets and capable of handling various downstream tasks related to seismogram analyzing, remains scarce. Currently, task-specific models are the norm for automation of seismogram processing. However, this could result in a lack of uniformity and efficiency in trained models. Meanwhile, in artificial intelligence fields such as Computer Vision (CV) and Natural Language Processing (NLP), various robust foundational network models have emerged, achieving significant success in their respective domains. Hence, this study posits the existence of a feature space suitable for multiple related seismological tasks, effectively representing seismic waveform information. This transformation allows the conversion of various seismic waveform analysis tasks into the problem of fine-tuning this feature space for specific tasks.

In this paper, a foundational model is designed, which functions as a feature extractor. Convolutional neural networks possess translational invariance and local induction bias, rendering them highly effective for extracting local features. However, they may lack the ability to establish long-term dependencies [24], which may be necessary for certain seismic waveform analysis tasks. To overcome this limitation, we introduce the global self-attention mechanism of the Transformer to capture interactions between global and local features [25]. Therefore, the design of this foundational model represents a hybrid network architecture combining Transformer [26] and convolution.

Furthermore, we observed challenges in the out-of-distribution generalization capabilities of existing models across different regions and seismic networks. Challenges may arise due to the differences in geology, varying epicentral distances, and other factors leading to differences in seismic waveform features. Inspired by the Short-Term Average to Long-Term Average (STA/LTA) algorithm [27], we designed Multi-Scale Mixed Convolution (MSMC) modules, Local Aggregation (LA) modules, and Multi-Path Transformer (MPT) modules to enhance the generalization capabilities of the model on different data distributions. Additionally, various output head modules are also designed, making this foundational network suitable for various tasks, including but not limited to earthquake detection, phase picking, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation.

In this paper, a novel neural network architecture called the Seismogram Transformer (SeisT) is proposed to address a key challenge in the application of deep learning to earthquake monitoring: the construction of a foundational network to enhance the efficiency and consistency of seismic waveform analysis tasks while ensuring generalization capabilities. We conducted model training on seismic data from mainland China and evaluated the performance of the model on seismic records from both mainland China and the Northwestern Pacific region of the United States, comparing it with state-of-the-art models for various tasks. The results demonstrate that SeisT consistently matches or exceeds the performance of existing approaches, particularly in terms of out-of-distribution generalization, for earthquake detection, phase picking, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation tasks. This model is of great reference to making significant advancements in seismological research and providing strong support for practical applications such as earthquake early warning and seismic monitoring.

2 Methods

2.1 Network architecture

The proposed novel network architecture (Fig. 1) consists of three main components: stem, body, and head. The stem performs initial processing on the input normalized seismograms by utilizing multi-path depthwise separable convolution and linear transformations. Through multiple layers of processing, it swiftly reduces the dimensionality of the input seismograms from l to $l/4$ along the time axis, aiming to alleviate the computational load of subsequent attention modules while obtaining fundamental feature representations. The body section comprises L stages, with each stage containing three parts: LA module, MSMC modules, and MPT modules. This structure enables the effective extraction of deep-level features. The head module is responsible for decoding the extracted features and outputting the target predictions. Multiple heads tailored for different tasks are designed, allowing the stem and body to share the same feature map across various tasks. The flexible modular design enables the model to be adapted to different types of tasks, including earthquake detection, seismic phase picking, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation. Due to the combined

Fig. 1 SeisT model architecture. The proposed model takes the normalized 3-component seismic waveform as input, preprocesses it through a stem module containing S stem layers, and then maps it to a high-dimensional feature space through L body modules. The i-th body module consists of a local aggregation module, C_i MSMC modules, and T_i MPT modules. It incorporates several fundamental modules, such as Group Convolution Block (GCB), Local Aggregation Multi-Head Attention (LAMHA), Depthwise Separable Convolution (DSConv), and Multi-Layer Perceptron (MLP). The model outputs vector representations specific to the given task through the head module.

structure of convolution and self-attention, the network architecture achieves independence from the length of seismograms during its construction. SeisT can handle seismograms of varying lengths without the need for predefined lengths during training and inference. Specifically, in the convolutional layers, it can automatically pad to match the convolution kernel size, and during up-sampling, it can dynamically calculate the up-sampling ratio for each stage based on the dimensions of the input seismograms. For regression and classification tasks, global average pooling is utilized to integrate feature maps of different shapes.

The stem and body modules collaboratively work to map the seismic waveforms into a feature space with compressed time dimensions and increased channel depth. This contributes to capturing a broader range of temporal information and the complex structures and patterns within the input data. Within the body, the multi-head self-attention structure plays a crucial role, which efficiently captures long-range dependencies, thus mitigating the limitations of receptive fields in convolutional networks. In the self-attention mechanism, for seismograms with a length of L and a channel size of C, the algorithm complexity is exceedingly high, reaching $\mathcal{O}(L^2 C)$. Therefore, the combination of local aggregation and self-attention mechanisms reduces the dimensions of key and value, significantly reducing computational overhead without decreasing the output dimensionality. At the beginning of each stage, there is a local aggregation module responsible for down-sampling the temporal dimension of the feature maps, performing local fusion on the feature map with a specified window size for each stage. These down-sampled features undergo feature extraction through a series of MSMC modules, resulting in more advanced and rich feature representations. Subsequently, these feature maps are inputted into one or more concatenated MPT modules, allowing deep-level features to be fused globally through attention mechanisms. Each module includes multiple residual connections and normalization layers, which facilitate the convergence of deep neural networks and reduce training difficulties.

For various tasks, three types of head structures have been designed. The first type of head is designed for earthquake detection and phase picking, composed of multiple consecutive up-sampling stages. Each stage includes linear interpolation and convolution operations to restore multiple down-sampled feature

map sizes corresponding to the stem and body. Then, the convolution modules and the sigmoid function are employed to integrate the output, mapping high-level feature information into three probability sequences related to the existence of seismic signals at each time point, phase-P, and phase-S. The second type of head is designed for classification tasks utilized for first-motion polarity classification. It incorporates global average pooling to aggregate features and reduces computational complexity, enabling the head to adapt to input feature maps with different sizes in the time dimension. Subsequently, a linear transformation and the softmax function are applied to output a vector of dimension 2. The third type of head is designed for regression tasks. Similar to the aforementioned second type (classification head), it shares commonalities but differs in that the final output undergoes sigmoid scaling instead of utilizing softmax normalization. This type of head is employed for regression tasks such as back-azimuth estimation, magnitude estimation, and epicentral distance estimation.

The SeisT designed in this paper is classified into three sizes based on its number of layers: Small (S), Medium (M), and Large (L). This categorization is intended to meet the requirements of different application scenarios and varying training data volumes. Including the head, the smallest network has a total parameter count of only 98k, while the largest network does not exceed 670k. The architecture of these networks draws inspiration from several visual and natural language processing models [28–32] and is designed based on domain-specific knowledge in seismology. The selection of model structure and hyperparameters is informed by extensive experimentation with numerous prototype networks (Supporting Table S1).

2.2 Network design

During seismic events, seismic signals are typically captured by multiple stations at varying distances and of different types. Due to differences in propagation paths and travel times, the signal characteristics received by various stations in the seismic network exhibit certain disparities. To address the variations in the temporal dimension, we introduced grouped convolutions with multiple different-sized kernels to allow the network to flexibly extract local signal features under various receptive field conditions. This helps mitigate the information loss caused by a single receptive field, allowing the model to handle the sample differences present in the temporal dimension. Convolutional neural networks possess a strong inherent bias for efficiently extracting local signal features. However, for tasks such as earthquake detection that require a focus on global features, convolutional networks may constrain the interrelation of global features. Transformer modules have been proven to be an effective architecture for seismic signal processing and modeling [8]. Its self-attention mechanism promotes the fusion of global features. In the original Transformer architecture, each token represents a single word. When applied to the seismograms, the information content of an individual sampling point is minimal. In processing seismic waveforms, one token corresponds to multiple samples at a given moment in time. Therefore, we designed the local aggregation module to represent local features by aggregating information from multiple nearby samples. This allows for interactions between local features, replacing point-to-point feature interactions. When the LA module is applied to the Transformer layer, it can reduce the dimensionality of key and value in the multi-head self-attention mechanism, thus reducing time complexity.

To further reduce information loss and computational overhead, the MPT module was designed to parallelize the Transformer and convolution operations effectively. This module reduces the computational load of multi-head self-attention and supplements global fusion features with the local feature extraction capabilities of convolutional neural networks. The foundational model employs the stem block for preprocessing seismic signals to adapt them for subsequent deep feature extraction. Various types of output heads were designed for different tasks (Supporting Figure S1), enabling the flexible application of this foundational model to multiple tasks related to seismic signals.

2.3 Multi-scaled mixed convolution

In order to more efficiently model seismic waveforms and features with the existing convolutional modules, we designed an efficient and adaptable convolutional module called the Multi-Scale Mixed Convolution. This module enhances the feature generalization ability of representation and accelerates inference speed. With inspiration from various modules such as Inception [31] and Xception [33], we project the input shallow-level features through linear projection into multiple feature spaces, denoted as $\{\mathcal{X}_i \in \mathbb{R}^{d_i \times t} \mid i = 1, 2, \cdots, s\}$, where d_i represents the number of output channels of the i-th sub-convolution module; t represents the length of the time dimension; s represents the total number of sub-convolution modules. This approach enables the model to consider feature space information from diverse scales and positions, promoting the effective learning of representations. The proposed MSMC is defined as follows:

$$\text{MSMC}(\mathbf{X}) = \text{BN}(\mathcal{C}(\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_s)), \tag{1}$$

where,

$$\mathbf{S}_i = \text{GCB}(\text{BN}(\mathbf{X}^\top \mathbf{W}_i + \mathbf{b}_i)). \tag{2}$$

Here, $\mathcal{C}(\cdot)$ is the concatenation operation; X represents the input to the MSMC. We perform concatenation along the channel axis to merge all features and normalize them through BatchNorm, denoted as $\text{BN}(\mathbf{X}) = \frac{\mathbf{X} - \mathbb{E}(\mathbf{X})}{\sqrt{\text{Var}(\mathbf{X}) + \epsilon}} \cdot \gamma + \beta$. γ and β are learnable affine transformation parameters, where ϵ is a small constant. $\mathbf{S}_i$ represents the output of the i-th sub-convolution module; $\mathbf{W}_h \in \mathbb{R}^{c \times d_h}$ is the learnable parameter matrix for the i-th linear transformation; and $\mathbf{b}_i \in \mathbb{R}^{d_i}$ is the bias. $\text{GCB}(\cdot)$ denotes the operation of the proposed grouped convolution block (Supporting Figure S2), which is computed as follows:

$$\text{GCB}(\mathbf{X}) = \mathbf{X}_r + \text{GELU}(\text{BN}(\mathbf{X}_r)^\top \mathbf{W}_1 + \mathbf{b}_1)^\top \mathbf{W}_2 + \mathbf{b}_2, \tag{3}$$

where,

$$\mathbf{X}_r = \mathbf{X}_o + \mathbf{X}, \tag{4}$$

$$\mathbf{X}_o = \text{GELU}(\text{BN}(\Psi_g(\mathbf{X})))^\top \mathbf{W}_o + \mathbf{b}_o. \tag{5}$$

Here, $\text{GELU}(x) = 0.5x(1 + \tanh(\sqrt{2/\pi}(x + 0.044715x^3)))$; $\Psi_g(\cdot)$ represents the standard grouped convolution operation. Experimental results from ablation studies indicate that the MSMC outperforms standard convolutional operations.

2.4 Multi-path Transformer

Transformer has demonstrated outstanding performance in addressing the problem of long-range dependencies [26]. However, its computational complexity reaches $\mathcal{O}(N^2)$, which is not feasible for long sequences of seismic wave inputs. Considering the efficiency and generalization capabilities of large neural networks, we designed the MPT module to incorporate local aggregation and convolution while maintaining a sparse structure. This sparse architecture helps reduce computational demands and mitigate overfitting issues. The similarity matrices in the Transformer have been proven to be low-rank matrices [34], suggesting that information from similarity matrices can be obtained by a small number of largest singular values. Hence, key and value can be mapped to lower-dimensional subspaces to reduce computational costs.

Fig. 2 Multi-path Transformer architecture. First, the input tensor is projected into two distinct lower-dimensional subspaces. Subsequently, it is separately fed into the Transformer with local aggregation and the grouped convolution block parallel architecture. Finally, feature fusion is achieved through concatenation with a multi-layer perceptron.

The MPT module employs a parallel structure design (Fig. 2), where the input tensor is linearly projected into two distinct lower-dimensional subspaces. These subspaces then enter a self-attention module with local perception and a grouped convolution module separately. Some prior studies utilized operations such as convolution and pooling to further reduce computational complexity [35–37]. In our MPT module, we utilized the proposed local aggregation module, which aggregated the key and value of the multi-head self-attention through local aggregation, enabling each token to incorporate contextual relationships in the temporal dimension.

Given the kernel size k_a for local aggregation, for a tensor with shape $c \times l$, the transformation through local aggregation results in a shape $c \times \lceil l/k_a \rceil$, thereby reducing the computational cost of the

similarity matrix. It is then restored to its original shape when combined with a query through dot-product attention. The computation process for multi-head self-attention with local aggregation can be represented as:

$$\text{LAMHA}(\mathbf{X}) = \mathcal{C}(\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_h)^\top \mathbf{W}_{ao} + \mathbf{b}_{ao}, \tag{6}$$

where,

$$\mathbf{H}_i = \text{softmax}(\frac{\mathbf{Q}_i^\top \mathbf{K}_i}{\sqrt{\mathbf{d}_h}})\mathbf{V}_i, \tag{7}$$

$$\mathbf{Q}_i = \mathbf{X}^\top \mathbf{W}_i^Q + \mathbf{b}_i^Q, \tag{8}$$

$$\mathbf{K}_i = \text{BN}(\text{LA}(\mathbf{X}, k_a))^\top \mathbf{W}_i^K + \mathbf{b}_i^K, \tag{9}$$

$$\mathbf{V}_i = \text{BN}(\text{LA}(\mathbf{X}, k_a))^\top \mathbf{W}_i^V + \mathbf{b}_i^V, \tag{10}$$

where $\mathcal{C}(\cdot)$ is the concatenation operation; $\mathbf{H}_i$ represents the output of the i-th head; $\mathbf{W}_{ao} \in \mathbb{R}^{d_a \times d_a}$ and $\mathbf{b}_{ao} \in \mathbb{R}^{d_a}$ are the learnable parameter matrices and bias parameter vectors for the linear projection layer before the output; $\mathbf{Q}_i$, $\mathbf{K}_i$ and $\mathbf{V}_i$ respectively represent the query, key and value vectors of the i-th head; $\mathbf{W}_i^Q \in \mathbb{R}^{d_a \times d_h}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_a \times d_h}$ and $\mathbf{W}_i^V \in \mathbb{R}^{d_a \times d_h}$ are the learnable parameter matrices of the i-th head respectively; $\mathbf{b}_i^Q \in \mathbb{R}^{d_h}$, $\mathbf{b}_i^K \in \mathbb{R}^{d_h}$ and $\mathbf{b}_i^V \in \mathbb{R}^{d_h}$ are the learnable bias parameter vectors of the i-th head, respectively; and $\text{LA}(\cdot, \cdot)$ represents the local aggregation operation.

The grouped convolution block includes multiple modules such as the convolution layer, linear projection layer, and MLP, which are used to assist Transformer feature extraction. The last two parts of features are spliced and an MLP is used to fuse the features, which is expressed as follows:

$$\text{MPT}(\mathbf{X}) = \mathbf{X}_o + \text{GELU}(\mathbf{X}_o^\top \mathbf{W}_{o1})^\top \mathbf{W}_{o2} + \mathbf{b}_{o2}, \tag{11}$$

where,

$$\mathbf{X}_o = \text{BN}(\mathcal{C}(\mathbf{X}_a + \text{LAMHA}(\mathbf{X}_a), \mathbf{X}_c + \text{GCB}(\mathbf{X}_c))), \tag{12}$$

$$\mathbf{X}_a = \text{BN}(\mathbf{X}^\top \mathbf{W}_a + \mathbf{b}_a), \tag{13}$$

$$\mathbf{X}_c = \text{BN}(\mathbf{X}^\top \mathbf{W}_c + \mathbf{b}_c). \tag{14}$$

Here, $\mathcal{C}(\cdot)$ is the concatenation operation; $\mathbf{W}_a \in \mathbb{R}^{c \times d_a}$, $\mathbf{b}_a \in \mathbb{R}^{d_a}$, $\mathbf{W}_c \in \mathbb{R}^{c \times d_c}$, $\mathbf{b}_c \in \mathbb{R}^{d_c}$ are learnable parameters for linear transformations before concatenation.

2.5 Local aggregation

In modeling seismic waveforms, we pay closer attention to their temporal variations, recognizing that a single time point cannot adequately capture the rich information contained in seismic data. Therefore, we proposed a concise and efficient structure called the local aggregation module (Fig. 3) to obtain feature fusion representations within a specified time step in the time dimension.

Fig. 3 Local aggregation module. This module reduces the size in the time dimension, fuses local features through two transformation functions, increases channel depth through a linear projection layer, and uses BatchNorm for normalization.

In this study, f_1 and f_2 are implemented using average pooling and max pooling, respectively. Max pooling selectively retains the maximum value among feature points within a local neighborhood, preserving fine-grained details while reducing the overall signal strength of high-frequency components. This approach is particularly effective for segments with abrupt changes in sequences, as it places greater emphasis on low-frequency feature components. [38, 39]. However, since max pooling only focuses on the maximum value, it may overlook other important information. In contrast, average pooling prioritizes the preservation of background information and maintains the overall distribution of frequency components. However, due to its equal weighting of all feature points, average pooling may introduce some blurring of detailed information, especially in applications where maintaining fine-grained details is crucial.

Hence, we proposed the local aggregation Module, which could simultaneously retain background information and enhance detailed features. This module effectively reduces the temporal dimension of features while increasing the feature depth. The computation process for this module can be represented as:

$$\text{LA}(\mathbf{X}, k) = \text{BN}([\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n]^\top \mathbf{W}_a + \mathbf{b}_a), \tag{15}$$

where $\mathbf{W}_a \in \mathbb{R}^{\mathbf{c}_d \times \mathbf{c}_{d+1}}$, $\mathbf{b}_a \in \mathbb{R}^{c_{d+1}}$; and c_d represents the number of channels of the d-th stage; $\mathbf{x}_i$ is a column vector representing the aggregated output within a specified time step. The dimension reduction in the time dimension essentially involves weighted fusion with a strong inductive bias. The weighting method and the calculation of weights w_j for time points j are as follows:

$$\mathbf{x}_i = \sum_{j=i \cdot k}^{i \cdot k + k - 1} w_j \mathbf{X}_{:,j}, \tag{16}$$

where,

$$w_j = \begin{cases} (1+k)/k & j \in M \\ 1/k & \text{otherwise,} \end{cases} \tag{17}$$

$$M = \{m_j \mid m_i = \text{argmax}(\mathbf{X}_{:,i \cdot k : i \cdot k + k})\}. \tag{18}$$

Here, $i = 0, 1, 2, \cdots, \lceil l/k \rceil - 1$.

3 Data and Results

3.1 Data and labeling

(a)

(b)

(c)

Fig. 4 Geographical distribution of DiTing and PNW datasets. (a) Comparison of seismic event distribution in both datasets. (b) Distribution of seismic events in the DiTing dataset (R1). (c) Visualization of seismic event distribution (R2), magnitude, and source depth in the PNW dataset.

We utilized the DiTing dataset [40] from the China region and the PNW dataset [41] from the Northwestern Pacific region of the United States. Both datasets comprise a substantial number of seismic events, including arrival times, first-motion polarity, and magnitude, among other labels. Notably, there are no overlapping seismic events in the two datasets. The geographical distribution of the two datasets is illustrated in Fig. 4. The distributions of signal-to-noise ratio (SNR), epicentral distance, and magnitude are shown in Supporting Figure S3.

We employed the DiTing dataset for training our neural network model. DiTing is a large-scale dataset obtained from the China Seismic Network (CSN), collected and organized by the Institute of Geophysics, China Earthquake Administration. It encompasses seismic events in mainland China and its neighboring regions spanning from 2013 to 2020. The magnitudes of these seismic events range from 0 to 7.7, with epicentral distances ranging from 0 to 330 kilometers. The waveform data have a length of 180 seconds and a sampling rate of 50 Hz. In this study, seismic event samples with complete labels were exclusively utilized to cater to the training requirements for multiple tasks. The dataset was randomly partitioned into a training set (80%), a validation set (10%), and a test set (10%). This encompassed approximately 277k three-component seismic waveform data samples from the China region.

To assess the cross-regional out-of-distribution generalization ability of the model, we conducted testing using the PNW dataset from the Pacific Northwest Seismic Network (PNSN). The "ComCat event" subset of the PNW dataset consists of data verified by analysts and sent to the United States Geological Survey (USGS), encompassing a total of 65,384 events spanning from 2002 to 2022. The seismic waveform data have a length of 150 seconds and a sampling rate of 100 Hz. To maintain consistency with the DiTing dataset labels, we retained samples with clear first-motion polarity and the ML magnitude type. This subset of data was used exclusively as a test set to evaluate the generalization capabilities of the model, thus not included in the training process.

During the training process, the earthquake detection task labels are represented using 0-1 vectors. These vectors start at the P-wave arrival time and end at a specified time denoted as $P_{arrival}+\lambda(S_{arrival}-P_{arrival})$, where in this study, $\lambda = 2$. As for the phase-picking task, a probability sequence is employed to indicate the probability of phase arrival times. We tested four different label forms: rectangular, triangular, Gaussian, and Sigmoid, to smooth the phase arrival positions [42]. In our hyperparameter selection process, the Gaussian label yielded lower loss and higher F1-Score, justifying its being applied in the final model. Under this label format, the probability of P-wave and S-wave arrival times was set to 1 at the manually marked positions and gradually decreased to 0 before and after that sampling point to follow a Gaussian distribution, with a total width of 0.5 seconds. For the first-motion polarity classification task, we utilized one-hot vectors to represent the first-motion polarity as upward and downward.

3.2 Evaluation

Various metrics were employed to evaluate the performance of the models on different tasks. The evaluation metrics chosen include Precision (Pr), Recall (Re), F1-Score (F1), Mean Error (Mean), Standard Deviation (Std.), Mean Absolute Error (MAE), and coefficient of determination (R^2), defined as follows:

$$\mathrm{Pr} = \frac{T_{\mathrm{p}}}{F_{\mathrm{p}} + T_{\mathrm{p}}}, \tag{19}$$

$$\mathrm{Re} = \frac{T_{\mathrm{p}}}{F_{\mathrm{n}} + T_{\mathrm{p}}}, \tag{20}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}, \tag{21}$$

$$\mathrm{Mean} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i), \tag{22}$$

$$\mathrm{Std} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}, \tag{23}$$

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{24}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y_i})^2}, \tag{25}$$

where T_{p} represents the number of true positives; F_{p} represents the number of false positives; F_{n} denotes the number of false negatives; N denotes the total number of samples; y_i represents the labels of sample i;

$\hat{y}_i$ represents the predictions of sample i; and $\overline{y}$ represents the mean value of the sample labels. Precision, recall, and F1-Score are common evaluation metrics in classification tasks. In the context of phase-picking tasks, samples with larger residuals are considered false positives. Accordingly, samples with residuals within the error tolerance $\delta < 0.1s$ were chosen as true positives in this study. The F1-Score is a balanced metric between precision and recall, providing a comprehensive evaluation of the performance of the model. MAE effectively measures the degree of offset between model predictions and sample true labels. Standard deviation indicates the stability of the predictive performance of the model. MAE is used to represent the magnitude of errors between model predictions and labels. The coefficient of determination (R^2) is used to measure the correlation between predicted values and true values in regression tasks.

3.3 Training details

When training SeisT, we initialized the linear and convolutional layers using truncated normal distribution [43], set BatchNorm weights to 1, and initialized all biases to 0. We utilized the cyclic learning rate scheduler [44] with the Adam optimizer [45], setting the minimum learning rate to 8×10^{-5} and the maximum learning rate to 1×10^{-3}. Periodic learning rate adjustments help prevent the model from getting stuck in local minima or saddle points. Early stopping was configured as follows: if the loss on the validation set did not decrease for 30 consecutive epochs, training was halted.

We employed data augmentation strategies on the training dataset, including adding random Gaussian white noise, time drift, introducing gaps, channel dropout, amplitude scaling, pre-emphasis, and generating noise with probabilities of 0.4, 0.4, 0.4, 0.4, 0.4, 0.97, and 0.05, respectively, in seismic waveforms. Regularization methods such as DropPath [46] and Dropout [47] were used with probabilities of 0.1, 0.2, and 0.3 for models S, M, and L, respectively.

We trained the models for seismic phase picking, earthquake detection, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation using the PyTorch framework on a single NVIDIA RTX-4090 GPU. The training time varies from 4 to 18 hours depending on model size and application task. To ensure a fair performance comparison among different models, the same data augmentation methods, optimizers, and learning rates were used for training models on the same tasks.

3.4 Results

3.4.1 Comparison with other methods

Seismograms of the "ENZ" (East, North, Vertical) components.

To comprehensively evaluate model performance, we conducted experiments across multiple tasks, including phase picking, earthquake detection, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation. To be exact, the top-performing models currently available were selected, including PhaseNet [11], EQTransformer [8], DiTingMotion [9], MagNet [16], and Baz-Network [18]. To ensure fairness and comparability of experiments, all these models were trained on the same training dataset and evaluated on the same test dataset. For each task, different models were trained using consistent training strategies and hyperparameters. It is worth emphasizing that we did not apply additional filtering to the test data to maintain the originality of the data. During the model training process, we employed various data augmentation techniques such as random cropping, time shifting, and amplitude scaling [48]. These techniques contributed to improving the convergence speed of the model, reducing the risk of overfitting, and improving the generalization of the trained model. We applied the same data augmentation strategies for both our study and the baseline models to ensure a fair comparison. Before inputting the data into the model, we normalized the waveforms to scale the amplitudes within the range suitable for the processing of the model.

In multiple tasks, the SeisT model has demonstrated superior performance, surpassing existing models or achieving comparable performance with them. The phase-P and phase-S picking tasks can be taken as examples. We compared SeisT with state-of-the-art deep learning models such as EQTransformer and PhaseNet (see Table 1). Six metrics, namely, Precision, Recall, F1-Score, Mean, Std, and MAE, were employed to evaluate model performance, where predictions with absolute errors smaller than the error tolerance were considered as true positives. An in-depth discussion and analysis of the relationship between model performance and error tolerance is also provided in Section 4. Due to significant feature differences between the PNW test dataset and the DiTing dataset, they do not adhere to the assumption of independent and identically distributed (i.i.d.) data, making the PNW test dataset highly effective in assessing the out-of-distribution generalization capability of the model. The test results on the i.i.d.

(a)

(b)

(c)

(d)

Fig. 5 Phase picking and earthquake detection examples. (a-d) are event examples from the test dataset, encompassing various SNRs, epicentral distances, and magnitudes. Specifically, (a) and (b) exhibit relatively high SNR values of 29 dB and 43 dB, respectively. The corresponding event of (a) has a magnitude of 1.7 and an epicentral distance of 37.5 km, and the corresponding event of (b) has a magnitude of 2.1 and an epicentral distance of 98 km. (c) and (d) exhibit lower SNR values of 4 dB and 1 dB, respectively. The corresponding event of (c) has a magnitude of 1.5 and an epicentral distance of 16 km, and the corresponding event of (d) has a magnitude of 5.0 with an epicentral distance of 216 km. Within the context of these examples, Z, N, and E represent the seismogram components. P and S represent the manual picking labels of P-wave and S-wave arrivals, respectively. $\hat{P}$ and $\hat{S}$ represent the predicted probabilities for P-wave and S-wave arrivals, respectively. $\hat{D}$ represents the probability of predicting the time from the P-wave arrival to the end of the S-wave coda.

dataset are shown in Supporting Table S3. On the PNW test dataset, compared with other models, the SeisT model demonstrates significant performance improvements. For the phase-P picking task, all three SeisT variants exhibit similar performance and achieve F1-Score improvements of 5.5% and 1.7% over PhaseNet and EQTransformer, respectively. Notably, the lightweight SeisT-S model utilizes only half of the PhaseNet parameters and one-third of the EQTransformer parameters. For the phase-S picking task, SeisT-M and SeisT-L similarly outperform existing models, achieving over 9% performance improvement over existing models. Even the lightweight SeisT-S achieves a performance improvement of over 5%. As is observed, the phase-P picking task typically exhibits higher precision than the phase-S picking task. This is because the P-waves coda partially overlaps with the initial arrivals of S-waves, making it more complex and challenging to accurately capture the features of S-waves. This in turn makes the phase-S picking task more challenging. In earthquake detection tasks, SeisT performs at a similar level to existing models across

Table 1 Model performance comparison

Task	Model	F1	Pr	Re	Mean	Std.	MAE	R^2
	PhaseNet	90.23	**96.91**	84.41	0.02	**0.03**	**0.02**	-
	EQTransformer	93.96	95.59	92.38	0.02	**0.03**	0.03	-
Phase-P	SeisT-S	95.66	96.70	94.64	0.02	**0.03**	**0.02**	-
	SeisT-M	95.63	96.85	94.44	0.02	**0.03**	**0.02**	-
	SeisT-L	**95.73**	96.62	**94.85**	**0.01**	**0.03**	**0.02**	-
	PhaseNet	57.54	69.06	49.32	**0.02**	**0.04**	**0.02**	-
	EQTransformer	58.40	62.10	55.12	0.03	**0.04**	0.03	-
Phase-S	SeisT-S	64.16	67.55	61.09	0.03	0.05	0.03	-
	SeisT-M	67.42	70.04	64.98	0.03	0.05	0.03	-
	SeisT-L	**67.98**	**70.83**	**65.35**	0.03	0.05	0.03	-
	EQTransformer	95.57	96.04	**95.10**	-	-	-	-
Detection	SeisT-S	95.13	**97.38**	92.99	-	-	-	-
	SeisT-M	95.43	96.43	94.45	-	-	-	-
	SeisT-L	**95.60**	96.71	94.52	-	-	-	-
	DitingMotion	94.26	94.40	94.14	-	-	-	-
First-Motion	SeisT-S	93.96	94.11	93.85	-	-	-	-
	SeisT-M	**94.70**	**94.79**	**94.64**	-	-	-	-
	SeisT-L	94.36	94.39	94.34	-	-	-	-
	MagNet	-	-	-	**0.00**	0.30	0.22	0.79
Magnitude	SeisT-S	-	-	-	**0.00**	0.24	0.18	0.86
	SeisT-M	-	-	-	**0.00**	**0.23**	**0.17**	**0.87**
	SeisT-L	-	-	-	0.01	0.24	0.18	0.86
	Baz-Network	-	-	-	**0.07**	69.47	43.89	0.55
Back-Azimuth	SeisT-S	-	-	-	-0.23	70.18	45.09	0.54
	SeisT-M	-	-	-	0.40	69.85	43.95	0.54
	SeisT-L	-	-	-	-0.14	69.56	43.96	**0.55**
	SeisT-S	-	-	-	-4.37	13.29	5.27	0.78
Distance	SeisT-M	-	-	-	**-3.81**	**12.23**	**4.80**	**0.81**
	SeisT-L	-	-	-	-4.01	12.72	5.03	0.80

F1, P, and R are F1-score, precision, and recall, respectively. Mean and Std are the mean and standard deviation of errors in seconds respectively. MAE is the mean absolute error. R^2 is the coefficient of determination. The training and testing of the magnitude estimation task were performed on the PNW dataset. The training and testing of the back-azimuth estimation task were performed on the DiTing dataset. The phase-picking, earthquake detection, first-motion polarity classification, and epicentral distance estimation tasks are trained using the DiTing dataset, and tested on the PNW dataset. Bold values represent the best performance.

various evaluation metrics. Some examples of time picking and earthquake detection are shown in Fig. 5. In the first-motion polarity classification task, SeisT achieves performance metrics similar to DiTingMotion, indicating their comparable performance levels. In the magnitude estimation task, metrics such as Mean, Std, MAE, and R^2 were used to evaluate model performance, and the results show that SeisT outperforms MagNet 2% in terms of determination coefficient on the test dataset. However, we also noted that the out-of-distribution generalization capability of the model is not outstanding in the magnitude estimation task. This may be due to the differences in seismic networks, geological features, magnitude types, and magnitude sizes in different regions. In the back-azimuth estimation task, the existing Baz-Network model requires the computation of covariance matrices, eigenvalues, and eigenvectors, followed by feature fusion through convolution and fully connected layers. This results in a parameter count exceeding 1050k, significantly higher than those of the SeisT series, namely, 98k, 312k, and 529k. Unexpectedly, SeisT achieves comparable or even superior performance with fewer parameters in the testing phase compared to Baz-Network. In the epicentral distance estimation task, all three models of SeisT achieved the R^2 of 0.98 on the DiTing test set. In cross-regional out-of-distribution generalization performance testing, the R^2 of three models reached 0.78, 0.81, and 0.80, respectively, and the MAE was only 5.27, 4.80, and 5.03, respectively.

3.4.2 Ablation study

We conducted ablation experiments to understand the contributions of the main modules in SeisT, as well as the effects of parameter count and training schedules. These experiments were conducted on the DiTing dataset, focusing on time picking and earthquake detection tasks. The following studies were performed: (a) Ablation study of MSMC module. (b) Ablation study of LA module. (c) Ablation study of MPT module. (d) Performance of SeisT variants with different parameter counts. (e) Performance of SeisT under different training schedules.

According to the ablation experimental results (Supporting Table S2), the MSMC module demonstrates a noticeable superiority over the regular convolution module, especially in the phase-P picking task, where its F1-Score improves by 1%, compared with the regular convolution module. The MSMC module, based on convolutional kernels of multiple sizes, is applied across different channels. Its superiority lies in its ability to extract rich multi-level features across multiple channels, effectively capturing underlying feature information and reducing the influence of irrelevant features on the results. Furthermore, The LA module is a simple yet effective structure. Experiments demonstrate that when employing a downsampling approach without the LA module, its performance decreases by 0.8% to 2.7% in earthquake detection and seismic phase picking tasks.

Compared to conventional self-attention mechanisms, the MPT module effectively mitigates computational burden in earthquake detection tasks while demonstrating notable performance enhancements. When considering the model without the attention mechanism path, the F1-Score for earthquake detection is only 88.98%. However, introducing the MPT module into the SeisT model leads to nearly an 8% performance improvement. This performance boost is attributed to the capability of the global self-attention mechanism to effectively capture long-range dependencies, which traditional convolution or recurrent neural networks struggle to capture. Furthermore, the attention mechanism brings about a 3% and 5% performance improvement for phase-P and phase-S picking tasks, respectively.

We tested models of three different sizes, S, M, and L, with the main differences being the feature map sizes and the number of layers within each stage. As expected, SeisT-L achieved the best performance on the DiTing dataset, while SeisT-S exhibited slightly lower performance metrics compared with SeisT-L, and SeisT-M fell in between the two. In the phase-P picking task, the F1-Score of SeisT-M is 0.02% higher than that of SeisT-L. Although this small performance difference may be negligible, it could be attributed to SeisT-L having a larger number of parameters, potentially causing it to slightly overfit the training data and thus achieving performance close to SeisT-M, which has only half the parameter count.

In a thorough exploration of hyperparameter choices, we observed that cyclic learning rate schedules [44], as opposed to linear decay or fixed learning rates, could significantly improve model performance. Cyclic learning rate schedules may help prevent the model from becoming trapped in local minima or saddle points, potentially leading to improved performance. As highlighted by Dauphin et al. [49], compared with local minima, saddle points hinder convergence more, and higher learning rates help to quickly escape saddle points. This is precisely the effect achieved by periodically increasing the learning rate.

4 Discussion

The SeisT model proposed in this paper demonstrates competitive or superior performance compared to existing state-of-the-art models in a variety of seismic monitoring tasks, including earthquake detection, seismic phase picking, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation. Its strong performance is primarily attributed to the model architecture design and data augmentation methods used during training.

The design of the MSMC and MPT plays a crucial role in extracting potential information from the seismograms. The self-attention mechanism with local aggregation empowers the neural network to focus on both local and global features simultaneously. The availability of the DiTing dataset provides ample training data, and the use of the PNW dataset allows for assessing cross-regional out-of-distribution generalization performance, particularly focusing on different regional distributions.

Within the SeisT model, the stem module is designed for preprocessing input seismograms. This module utilizes multi-path depth-wise separable convolutions and linear transformations to rapidly reduce the time dimension of seismograms. This significantly reduces the computational overhead for subsequent convolutional operations and attention mechanisms. Directly inputting the waveforms into multi-head self-attention layers would incur an unacceptable computation complexity of $\mathcal{O}(L^2)$. In the body module, the number of stages can be flexibly adjusted based on application requirements. In general, deeper

Fig. 6 Performance comparison of seismic phase picking among different models. (a) corresponds to the performance of phase-P picking, while (b) corresponds to the performance of phase-S picking. Samples with an absolute error less than the tolerance threshold (from 0.1s to 0.5s) are considered true positives. The models were trained on the DiTing dataset and evaluated on the PNW dataset.

networks tend to possess stronger non-linear fitting and feature representation capabilities, potentially leading to higher accuracy. However, deeper networks also come with increased computational complexity and a higher risk of overfitting. Therefore, the proposed models adopt a design with 4 stages, a structure utilized by various network models in the field [28–31]. This design effectively reduces dimensionality while extracting deep features.

Among numerous tasks, phase picking stands out as one of the most representative tasks. This is due to its requirement to provide probability values for each sample point in the corresponding waveform, which relies on the accurate extraction of seismic wave features. PhaseNet and EQTransformer are widely used phase-picking models. In our out-of-distribution generalization tests, we observed varying degrees of performance degradation in phase-P and phase-S picking for both models when using low error tolerance. Notably, S-wave performance witnessed a significant decline, with a decrease of over 20% in maximum F1-Score when the tolerance threshold was reduced from 0.2s to 0.1s. This decline can be attributed to the overlap of S-wave with P-wave arrivals, making accurate phase-S picking at low tolerance thresholds exceptionally challenging. Moreover, manual phase picking introduces a certain degree of subjectivity, potentially leading to poor model performance at low error tolerance settings.

Even under the demanding error tolerance threshold of 0.1s, compared with other picking models, SeisT exhibited significantly better performance. Specifically, SeisT-M outperformed EQTransformer by 1.7% and 9.0% in P-wave and S-wave arrival time picking, respectively, on the out-of-distribution test set. SeisT consistently showed superior performance across different tolerance thresholds (ranging from 0.1s to 0.5s) on the out-of-distribution test set, as demonstrated in Fig. 6. This suggests that the SeisT model adapts well to seismic waveforms monitored by different networks in various regions, exhibiting robust out-of-distribution generalization capabilities. Such capabilities significantly reduce the transfer learning cost when the model is to be applied in practice.

The ability to pick arrivals is correlated with the SNR, epicentral distance, and magnitude of seismic waveforms. As shown in Fig. 7, the relationship between the number of P-wave and S-wave picks in different models and the SNR, epicentral distance, and magnitude is depicted. While this distribution trend is determined by the distribution of the data samples themselves, the disparities observed across different models can still reflect variations in model performance. The application of phase-picking models on low signal-to-noise ratio waveforms has long been a challenge. Compared with the baseline model, the SeisT series models exhibit noticeable performance improvements in low signal-to-noise ratio conditions, especially in the signal-to-noise ratio range of 0-10 dB, demonstrating significant differences. Under different epicentral distances, substantial differences are mainly manifest in the phase-S picking of seismic events

13

Fig. 7 Comparison of phase picking performance for earthquake events with different SNRs, epicentral distances, and magnitudes. (a) and (b) represent the performance distribution of various phase-picking models under different SNR conditions for phase-P picking and phase-S picking, respectively. (c) and (d) represent the performance distribution of various phase-picking models under different epicentral distance conditions for phase-P picking and phase-S picking, respectively. (e) and (f) represent the performance distribution of various phase-picking models under different magnitude levels for phase-P picking and phase-S picking, respectively. The test data utilized in this analysis are derived from the PNW dataset.

within the 0-50 km range. Different seismic magnitudes can exhibit variations in signal characteristics, with smaller magnitude events implying weaker signal strength, potentially increasing the difficulty of signal analysis. In the low-magnitude range, SeisT continues to outperform other models, indicating its higher sensitivity and ability to ensure picking performance even for weaker signals. Fig. 8 illustrates the residual distribution of different models in phase-P and phase-S picking tasks. In the vicinity of zero, the SeisT series performs the best, with PhaseNet slightly outperforming EQTransformer in phase-P picking and demonstrating similar performance in phase-S picking. One possible reason for this is that extensive residual connections of PhaseNet allow it to retain a substantial amount of original fine-grained features, thus achieving better performance in the relatively straightforward phase-P picking task.

Fig. 8 Comparison of residual distributions between multiple models for automated phase picking and manual picking. (a) and (b) represent the residual distribution of P-wave and S-wave arrival time picking, respectively. The test data used are from the PNW dataset.

Fig. 9 The performance of SeisT in the magnitude estimation task. (a) represents the performance of SeisT-L on the DiTing test dataset, achieving R^2 of 0.94. (b) denotes the performance of SeisT-L on the PNW test dataset, with the R^2 value of 0.86.

First-motion polarity classification is a classification task that primarily relies on waveform characteristics in the immediate aftermath of the P-wave arrival. Therefore, the model needs to input a segment of the waveform containing the P-wave arrival time and a period afterward. First-motion polarity is closely related to the trend of waveform changes following the P-wave arrival. DiTingMotion introduces an additional input channel, which is the amplitude difference between adjacent time points. This method can to some extent reduce the learning difficulty of waveform trends for the neural network. SeisT, on the other hand, does not adopt this approach. To maintain simplicity and computational efficiency, SeisT utilizes the normalized three-component seismic waveforms as input for the first-motion polarity classification task, without the need for any additional pre-computed data. Despite its straightforward approach, SeisT achieves comparable performance to DiTingMotion.

Fig. 10 t-SNE visualization of data sample feature representation. Randomly extracting an equal number of samples from the DiTing and PNW datasets, we visualized their feature representations using t-SNE. Here, U denotes that a small subset of samples from both datasets exhibits similarity in the high-dimensional complex feature space. S_1 and S_2 indicate that the majority of samples from the two datasets are distinctly separable in the high-dimensional complex feature space, signifying significant feature differences. C_1 and C_2 represent two feature dimensions.

In the seismic magnitude estimation task, the data distribution unexpectedly influences the models. Unlike several other tasks, seismic magnitude estimation requires training the model on data with similar distributions. For example, it is difficult to apply a model trained on data collected from the CSN to seismic waveforms collected from the PNSN due to differences in seismic network equipment. In this study, we separately trained SeisT on the DiTing and PNW datasets and evaluated them on their respective test sets. On the PNW test set, the R^2 values for SeisT-S, SeisT-M, and SeisT-L are 0.86, 0.87, and 0.86, respectively. Fig. 9 displays the relationship between the predicted and true magnitudes of SeisT-L on the DiTing and PNW test sets. MagNet performs similarly to SeisT on the DiTing dataset, with a coefficient of determination of 0.91 (see Supporting Table S3), and on the PNW dataset, it achieves the R^2 of 0.79, which is 0.07 to 0.08 lower than the SeisT series.

Due to the absence of official back-azimuth labels in the PNW dataset, the training and testing for the back-azimuth estimation task were solely performed using the DiTing dataset. In current applications, back-azimuth estimation often occurs after seismic source localization. Therefore, back-azimuth estimation based on deep neural networks helps rapidly determine back-azimuth angles, providing a valuable reference for subsequent work. Testing on the DiTing dataset reveals that Baz-Network and SeisT perform similarly in back-azimuth estimation, with coefficients of determination of approximately 0.54 and MAEs exceeding 40. While these results may not provide high-precision back-azimuth estimates, they still keep the average error within a sharp angle range. Using this model as a preliminary task for localization can contribute to the reliability of earthquake localization results.

The epicentral distance estimation is a typical regression task. In previous studies, it is often closely related to travel times and seismic source depth. In this study, we only utilize seismograms from a single station and epicentral distance labels for the regression task. The neural network, leveraging its powerful representation learning capability, effectively captures the correlation between seismograms and epicentral distance. In cross-regional out-of-distribution testing, the evaluation metrics R^2 for the three SeisT models are 0.78, 0.81, and 0.80, respectively. On the test set within the same region as the train set, all three models achieve the R^2 of 0.98 (see Supporting Table S3). It is worth noting that the task is sensitive to the sampling rate of input seismograms, necessitating the resampling of test data to match the sampling rate of the training data. In the experiments, we found that in the epicentral distance estimation task, when the sampling rates of training and testing data are inconsistent, unacceptable errors in results arise. We attribute this to the model not incorporating travel time information, leading the neural network to be incapable of explicitly learning wave velocities. From a traditional perspective of analyzing travel time differences, when the neural network learns the correlation between epicentral distance and the time differences of P-waves and S-waves, the unknown sampling rate of input seismograms causes the fitting target of the neural network to be the relationship between epicentral distance and the number of sampling points. Consequently, the sampling rate of input seismograms significantly influences the ability of the neural network to perceive the time differences between P-waves and S-waves.

As previously mentioned, to validate the out-of-distribution generalization ability of the models, we trained each of them on the DiTing dataset, which includes seismic events in China and neighboring

regions. Subsequently, we separately tested the models on the DiTing test set and the PNW dataset. Fig. 10 presents the feature distributions of samples from the DiTing and PNW datasets. These features are derived from the output of the SeisT feature extractor, namely the output of the body module. The t-distributed stochastic neighbor embedding (t-SNE) was applied to reduce the dimensionality of deep seismic waveform features. The U region in the figure indicates that a small portion of samples from both datasets exhibit some similarity in the high-dimensional complex feature space, making it challenging to distinguish them significantly. The samples in regions S_1 and S_2 highlight clear differences in the feature distributions for the majority of samples between the two datasets, which may be attributed to significant differences in geological conditions, crustal structures, and source characteristics among different regions. This suggests that the model can effectively differentiate samples from the two datasets using this feature representation, indicating that the model has learned a regional feature representation. Therefore, it can be inferred that there are significant feature differences between the two datasets, suggesting a lack of independence and identical distribution.

SeisT is designed as a foundational model for seismogram analysis, and its generalization performance on data from different regions has been experimentally validated. With its flexible architecture design, SeisT also holds the potential to address various challenges based on seismograms, such as event classification and seismic source depth estimation. By altering the input channel dimensions, it is also conceivable to apply SeisT to the analysis of seismograms from multiple stations, including earthquake location and focal mechanism inversion. However, our approach still has limitations, primarily manifested in tasks sensitive to sampling rates and amplitudes. The generalization capacity for seismograms with varying sampling rates and amplitudes requires further enhancement. We encourage the exploration of more advanced representation learning methods to endow the model with broader out-of-distribution generalization capabilities.

5 Conclusion

In this study, we employed a network model with a hybrid design of Transformers and convolution, serving as a foundational model. It can be applied to a variety of tasks, including but not limited to earthquake detection, phase picking, first-motion polarity classification, magnitude estimation, back-azimuth estimation, and epicentral distance estimation in the fields of seismology. We trained the network model on the DiTing dataset, consisting of seismic data from China and its surrounding regions, using approximately 221k training samples, and evaluated its out-of-distribution generalization performance using 90k test samples of the PNW dataset from the Northwestern Pacific region of the United States. The results indicate that this model exhibits superior out-of-distribution generalization capabilities compared to several existing baseline models. This model holds substantial potential for diverse applications in seismic monitoring, thanks to its robust network architecture, which can play a pivotal role in advancing work related to seismic waveform analysis.

Code and Data Available

Our source code and model are available at https://github.com/senli1073/SeisT and can reproduce the results presented in this paper. The DiTing [40] dataset was used for training, validation, and testing and can be obtained at https://www.doi.org/10.12080/nedc.11.ds.2022.0002. The PNW [41] dataset was exclusively used for testing the out-of-distribution generalization ability and can be downloaded at https://github.com/niyiyu/PNW-ML. Maps and figures in this paper were generated using PyGMT [50] and Matplotlib [51].

6 Acknowledgements

7 Competing interests

The authors declare no competing interests.

References

[1] Ross, Z.E., Meier, M., Hauksson, E., Heaton, T.H.: Generalized seismic phase detection with deep learning. Bulletin of the Seismological Society of America **108**, 2894–2901 (2018) https://doi.org/10.1785/0120180080

[2] Perol, T., Gharbi, M., Denolle, M.: Convolutional neural network for earthquake detection and location. Science Advances **4**(2), 1700578 (2018)

[3] Mousavi, S.M., Beroza, G.C.: Deep-learning seismology. Science **377**, 4470 (2022) https://doi.org/10.1126/science.abm4470

[4] Mohammadigheymasi, H., Shi, P., Tavakolizadeh, N., Xiao, Z., Mousavi, S.M., Matias, L., Pourvahab, M., Fernandes, R.: Ipiml: A deep-scan earthquake detection and location workflow integrating pair-input deep learning model and migration location method. IEEE Transactions on Geoscience and Remote Sensing (2023)

[5] Wang, K., Hu, T., Zhao, B.: An unsupervised deep neural network approach based on ensemble learning to suppress seismic surface-related multiples. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)

[6] Mousavi, S.M., Beroza, G.C.: Machine learning in earthquake seismology. Annual Review of Earth and Planetary Sciences **51**(1), 105–129 (2023) https://doi.org/10.1146/annurev-earth-071822-100323

[7] Mousavi, S.M., Zhu, W., Sheng, Y., Beroza, G.C.: Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection. Sci Rep **9**, 10267 (2019) https://doi.org/10.1038/s41598-019-45748-1

[8] Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y., Beroza, G.C.: Earthquake transformer-an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature Communications **11**, 3952 (2020) https://doi.org/10.1038/s41467-020-17591-w

[9] Xiao, Z., Wang, J., Liu, C., Li, J., Zhao, L., Yao, Z.: Siamese earthquake transformer: A pair-input deep-learning model for earthquake detection and phase picking on a seismic array. Journal of Geophysical Research: Solid Earth **126**(5) (2021) https://doi.org/10.1029/2020jb021444

[10] Liao, W.-Y., Lee, E.-J., Chen, D.-Y., Chen, P., Mu, D., Wu, Y.-M.: Red-pan: Real-time earthquake detection and phase-picking with multitask attention network. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–11 (2022)

[11] Zhu, W., Beroza, G.C.: Phasenet: A deep-neural-network-based seismic arrival time picking method. Geophysical Journal International **216**, 261–273 (2018) https://doi.org/10.1093/gji/ggy423

[12] Yu, Z., Wang, W.: Lppn: A lightweight network for fast phase picking. Seismological Research Letters **93**(5), 2834–2846 (2022) https://doi.org/10.1785/0220210309

[13] Zhao, M., Xiao, Z., Zhang, M., Yang, Y., Tang, L., Chen, S.: Ditingmotion: A deep-learning first-motion-polarity classifier and its application to focal mechanism inversion. Frontiers in Earth Science **11**, 1103914 (2023) https://doi.org/10.3389/feart.2023.1103914

[14] Ross, Z.E., Meier, M., Hauksson, E.: P wave arrival picking and first-motion polarity determination with deep learning. Journal of Geophysical Research: Solid Earth **123**, 5120–5129 (2018) https://doi.org/10.1029/2017jb015251

[15] Chakraborty, M., Cartaya, C.Q., Li, W., Faber, J., Rümpker, G., Stoecker, H., Srivastava, N.: Polarcap – a deep learning approach for first motion polarity classification of earthquake waveforms. Artificial Intelligence in Geosciences **3**, 46–52 (2022) https://doi.org/10.1016/j.aiig.2022.08.001

[16] Mousavi, S.M., Beroza, G.C.: A machine-learning approach for earthquake magnitude estimation. Geophysical Research Letters **47**(1) (2020) https://doi.org/10.1029/2019gl085976

[17] Wang, Z., Liao, J., Wwang, Y., Wei, D., Zhao, D.: A fast magnitude estimation method based on deep convolutional neural networks. Chinese Journal of Geophysics **66**(1), 272–288 (2023)

[18] Mousavi, S.M., Beroza, G.C.: Bayesian-deep-learning estimation of earthquake location from single-station observations. IEEE Transactions on Geoscience and Remote Sensing **58**(11), 8211–8224 (2020) https://doi.org/10.1109/tgrs.2020.2988770

[19] Andreas, K., Erik, B.M.: Arraynet: A combined seismic phase classification and back-azimuth regression neural network for array processing pipelines. Bulletin of the Seismological Society of America (2023) https://doi.org/10.1785/0120230056

[20] Ristea, N.-C., Radoi, A.: Complex neural networks for estimating epicentral distance, depth, and magnitude of seismic waves. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022)

[21] Si, X., Wu, X., Sheng, H., Zhu, J., Li, Z.: SeisCLIP: A seismology foundation model pre-trained by multi-modal data for multi-purpose seismic feature extraction (2023)

[22] Sheng, H., Wu, X., Si, X., Li, J., Zhang, S., Duan, X.: Seismic Foundation Model (SFM): a new generation deep learning model in geophysics (2023)

[23] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the Opportunities and Risks of Foundation Models (2022)

[24] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

[25] Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Conference and Workshop on Neural Information Processing Systems (2017)

[27] Allen, R.V.: Automatic earthquake recognition and timing from single traces. Bulletin of the Seismological Society of America **68**(5), 1521–1532 (1978)

[28] Ze, L., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision (2021)

[29] Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., Yu, G., Shen, C.: Topformer: Token pyramid transformer for mobile semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2022)

[30] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S.,

Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. Proceedings of the IEEE conference on computer vision and pattern recognition, 6877–6886 (2020)

[31] Si, C., Yu, W., Zhou, P., Zhou, Y., Wang, X., Yan, S.: Inception transformer. In: Conference and Workshop on Neural Information Processing Systems (2022)

[32] Wang, Z., Cun, X., Bao, J., Liu, J.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2022)

[33] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

[34] Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. Preprint at https://arxiv.org/pdf/2006.04768 (2020)

[35] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

[36] Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)

[37] Li, J., Xia, X., Li, W., Li, H., Wang, X., Xiao, X., Wang, R., Zheng, M., Pan, X.: Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. Preprint at https://arxiv.org/pdf/2207.05501 (2022)

[38] Hoshen, Y., Weiss, R.J., Wilson, K.W.: Speech acoustic modeling from raw multichannel waveforms. In: International Conference on Acoustics, Speech and Signal Processing (2015)

[39] Sainath, T.N., Weiss, R.J., Senior, A.W., Wilson, K.W., Vinyals, O.: Learning the speech front-end with raw waveform cldnns. In: Interspeech (2015)

[40] Zhao, M., Xiao, Z., Chen, S., Fang, L.: Diting: A large-scale chinese seismic benchmark dataset for artificial intelligence in seismology. Earthquake Science **36**(2), 84–94 (2023) https://doi.org/10.1016/j.eqs.2022.01.022

[41] Ni, Y., Hutko, A., Skene, F., Denolle, M., Malone, S., Bodin, P., Hartog, R., Wright, A.: Curated pacific northwest ai-ready seismic dataset. Seismica **2** (2023) https://doi.org/10.26443/seismica.v2i1.368

[42] Nguyen, Q., Valizadegan, H., Hauskrecht, M.: Learning classification models with soft-label information. Journal of the American Medical Informatics Association **21**(3), 501–508 (2014)

[43] Jinho, C., Byung, R.C., Julia, L.S.: Rethinking the truncated normal distribution. International Journal of Experimental Design and Process Optimisation **3**, 327 (2013) https://doi.org/10.1504/IJEDPO.2013.059667

[44] Smith, L.N.: Cyclical learning rates for training neural networks. In: Winter Conference on Applications of Computer Vision (WACV) (2017)

[45] Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)

[46] Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European Conference on Computer Vision (2016)

[47] Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014) https://doi.org/10.5555/2627435.2670313

[48] Zhu, W., Mousavi, S.M., Beroza, G.C.: Seismic signal augmentation to improve generalization of deep neural networks. Advances in geophysics **61**, 151–177 (2020) https://doi.org/10.1016/bs.agph.2020.07.003

[49] Yann, D., Harm, d.V., B., Y.: Equilibrated adaptive learning rates for non-convex optimization. In: Conference and Workshop on Neural Information Processing Systems (2015)

[50] Uieda, L., Tian, D., Leong, W.J., Toney, L., Schlitzer, W., Grund, M., Newton, D., Ziebarth, M., Jones, M., Wessel, P.: PyGMT: A Python interface for the generic mapping tools (2021). https://doi.org/10.5281/ZENODO.4522136

[51] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in science & engineering **9**(03), 90–95 (2007)