

ARRU Phase Picker: Attention Recurrent-Residual U-Net for Picking Seismic *P*- and *S*-Phase Arrivals

Wu-Yu Liao¹, En-Jui Lee^{*1} , Dawei Mu² , Po Chen³ , and Ruey-Juin Rau¹ 

Abstract



Seismograms are convolution results between seismic sources and the media that seismic waves propagate through, and, therefore, the primary observations for studying seismic source parameters and the Earth interior. The routine earthquake location and travel-time tomography rely on accurate seismic phase picks (e.g., *P* and *S* arrivals). As data increase, reliable automated seismic phase-picking methods are needed to analyze data and provide timely earthquake information. However, most traditional auto-pickers suffer from low signal-to-noise ratio and usually require additional efforts to tune hyperparameters for each case. In this study, we proposed a deep-learning approach that adapted soft attention gates (AGs) and recurrent-residual convolution units (RRCUs) into the backbone U-Net for seismic phase picking. The attention mechanism was implemented to suppress responses from waveforms irrelevant to seismic phases, and the cooperating RRCUs further enhanced temporal connections of seismograms at multiple scales. We used numerous earthquake recordings in Taiwan with diverse focal mechanisms, wide depth, and magnitude distributions, to train and test our model. Setting the picking errors within 0.1 s and predicted probability over 0.5, the AG with recurrent-residual convolution unit (ARRU) phase picker achieved the F1 score of 98.62% for *P* arrivals and 95.16% for *S* arrivals, and picking rates were 96.72% for *P* waves and 90.07% for *S* waves. The ARRU phase picker also shown a great generalization capability, when handling unseen data. When applied the model trained with Taiwan data to the southern California data, the ARRU phase picker shown no cognitive downgrade. Comparing with manual picks, the arrival times determined by the ARRU phase picker shown a higher consistency, which had been evaluated by a set of repeating earthquakes. The arrival picks with less human error could benefit studies, such as earthquake location and seismic tomography.

Cite this article as Liao, W.-Y., E.-J. Lee, D. Mu, P. Chen, and R.-J. Rau (2021). ARRU Phase Picker: Attention Recurrent-Residual U-Net for Picking Seismic *P*- and *S*-Phase Arrivals, *Seismol. Res. Lett.* **92**, 2410–2428, doi: [10.1785/0220200382](https://doi.org/10.1785/0220200382).

Introduction

Seismograms generated by natural or man-made seismic sources and recorded as time-series data by seismometers are the most fundamental observations, both for studying the mechanisms of earthquakes (Aki and Richards, 2002) and for imaging the internal structure of the Earth (Iyer and Hirahara, 1993). On a typical seismogram, information about the seismic source and the Earth structure is unevenly distributed. Seismologists have long recognized that the dominant majority of useful information on a seismogram is concentrated on a handful of seismic phases (Storchak *et al.*, 2003, 2011), which are categories of waveform segments whose shapes and arrival times at the seismometer depend primarily upon the location and mechanism of the source, as well as the Earth structure sampled by the wavepath from the source to the seismometer. For regional seismic studies, such as those in Taiwan (Fig. 1a),

the two most important seismic phases on a seismogram are the first-arrived “*P*” (i.e., primary or pressure) wave (Aki and Richards, 2002) and the “*S*” (i.e., secondary or shear) wave (Aki and Richards, 2002) (Fig. 2a). A primary task in seismic data processing involves identifying the arrival time (i.e., time for the beginning) of each seismic phase on the seismogram. In the past few decades, rapid advances in seismic instrumentation, especially in earthquake-prone regions such as Taiwan

1. Department of Earth Sciences, National Cheng Kung University, Tainan, Taiwan,  <https://orcid.org/0000-0003-1545-1640> (EJL);  <https://orcid.org/0000-0003-1967-7612> (RJR); 2. National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, Illinois, U.S.A.,  <https://orcid.org/0000-0002-6354-6436> (DM); 3. Department of Geology and Geophysics, University of Wyoming, Laramie, Wyoming, U.S.A.,  <https://orcid.org/0000-0002-5148-9788> (PC)

*Corresponding author: rickli92@gmail.com

© Seismological Society of America

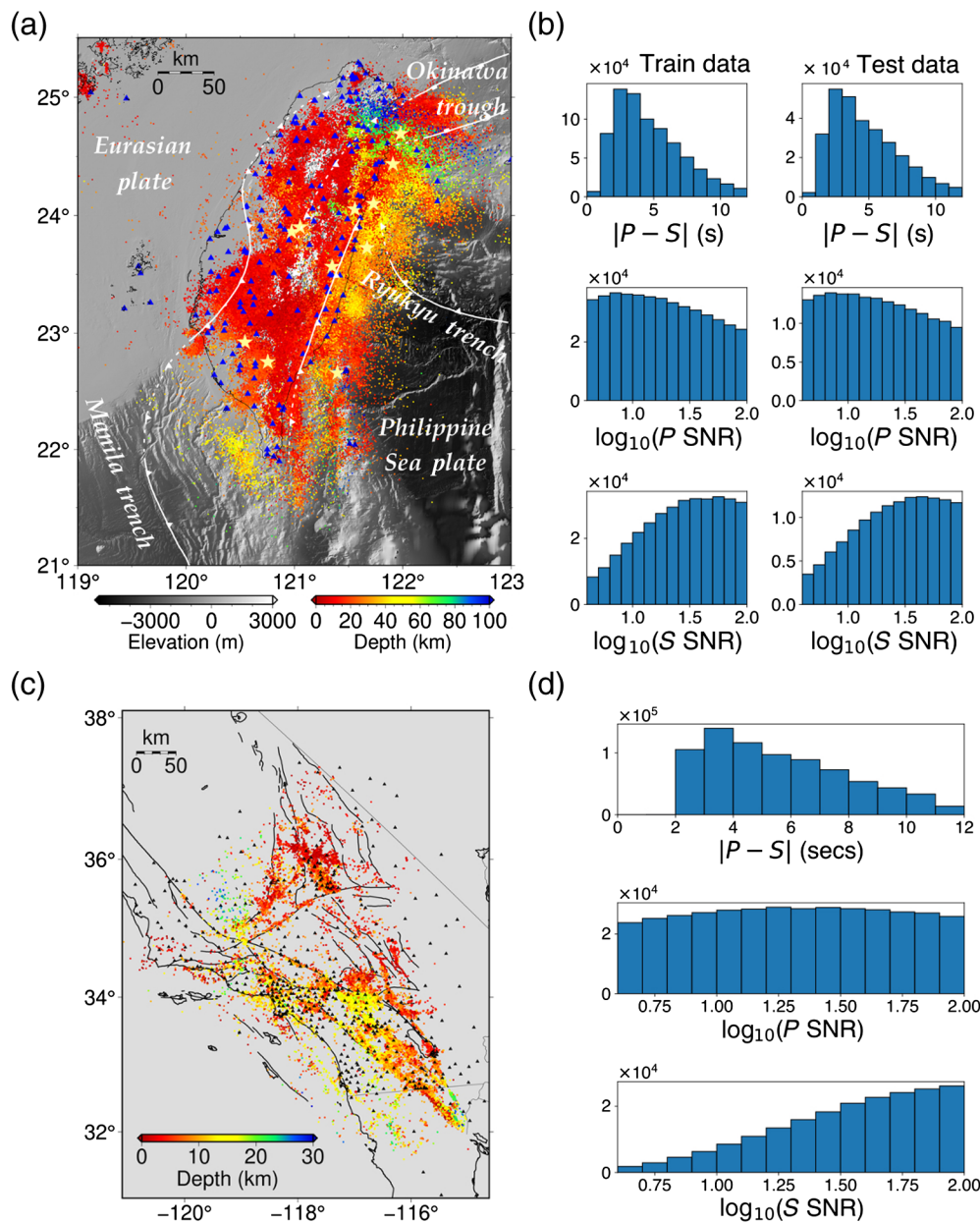


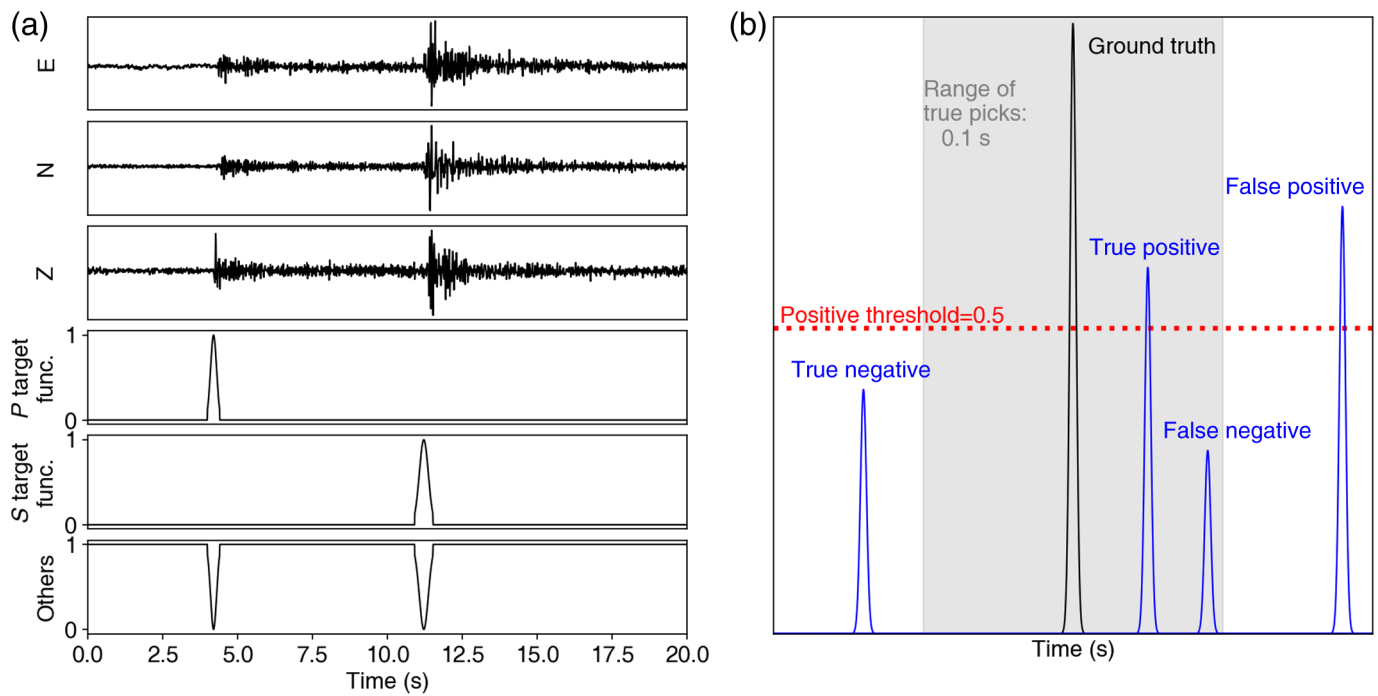
Figure 1. Basic information (P and S separations and signal-to-noise ratios [SNRs]) and spatial distribution of data used in this study. (a) The map shows the tectonic settings of Taiwan, and distributions of earthquakes and stations used in this study. Dots represent earthquake locations and colored by depths; blue triangles denote seismic stations; yellow stars show the earthquakes with M_L larger than 6.0. (b) Histograms show distributions of the P - and S -arrival residuals, and SNRs of P and S waves in our train and test datasets. (c) The map shows the distributions of earthquakes and stations of southern California (SC) used for examining model generalization in this study. (d) Histograms show distributions of the P - and S -arrival residuals, and SNRs of P and S waves in the SC dataset. The color version of this figure is available only in the electronic edition.

(Shin *et al.*, 1996, 2013), have led to an exponential growth in the volume of continuously recorded seismic data. It has become increasingly unrealistic to expect well-trained seismologists to manually mark the arrival times of seismic phases rapidly and reliably in such a massive, ever-increasing dataset. The only feasible approach forward is to automate the seismic

phase-picking task using computer algorithms that can offer comparable robustness and accuracy as well-trained seismologists.

Conventional algorithms for automatic seismic phase picking often rely upon detecting abrupt changes in the time series using statistical methods. A commonly adopted approach is to transform the seismogram into a time-dependent characteristic function that is more sensitive to abrupt changes, such as the ratio of short-term and long-term averages (STAs/LTAs; Allen, 1982), envelope functions (Baer and Kradolfer, 1987), autoregressive Akaike information criterion (AR-AIC; Sleeman and Van Eck, 1999), kurtosis (Saragiotis *et al.*, 2002; Baillard *et al.*, 2014), skewness (Nippres *et al.*, 2010; Ross and Ben-Zion, 2014), filtering (Lomax *et al.*, 2012), and particle-motion polarization (Jurkevics, 1988; Cichowicz, 1993; Baillard *et al.*, 2014). After applying some postprocessing steps on the characteristic function to reduce false detection rate, the arrival times of seismic phases are then picked automatically, based upon the temporal positions of local extrema on the characteristic function. Depending upon the signal-to-noise ratio (SNR) of the seismogram, the mechanism of the seismic source, as well as the physics of subsurface wave propagation, the arrival times of certain seismic phases may not always correspond to abrupt changes that can be

easily identified using the commonly adopted statistics. For seismograms with poor SNR, the picks generated by conventional algorithms still need to be reviewed and updated by seismologists before being used in other seismic applications. Moreover, some conventional phase-picking algorithms based on statistical methods often require considerable amount of human effort in



parameter tuning before being applied to different datasets, and inappropriate parameter settings can easily cause high false-positive results. On the other hand, the recently widely used template-matching algorithm (TMA) uses waveforms of *P* and/or *S* waves of confirmed earthquakes as templates for detecting earthquakes in continuous recordings (e.g., [Mu et al., 2017](#); [Ross et al., 2019](#); [Lee et al., 2020](#)). Because the TMA uses all the features, the complete waveforms, of *P* and/or *S* waves, the TMA can be very robust to avoid false positives. However, the TMA is limited to detecting events with similar focal mechanisms and very close to the template earthquake. Also, the TMA can be computationally expensive and requires an earthquake catalog for phase templates.

The choice of the characteristic function used in conventional phase-picking algorithms can be considered as statistic-based feature selection in the context of machine learning ([Guyon and Elisseeff, 2003](#)). The purpose of feature selection is to reduce the redundancy in the input data to those variables that have the strongest relationship with, therefore are most predictive of, the target. Unlike conventional phase-picking algorithms, in which the feature is prespecified by human based on statistic presumptions, the artificial neural network (ANN) allows automatic learning of the most predictive features through data-driven training. The features learned by the ANN are adapted to the specific task and the given dataset, thereby, offering the potential of improved accuracy and robustness. Early attempts of applying the ANN to automatic seismic phase picking ([Dai and MacBeth, 1995a,b, 1997](#); [Wang and Teng, 1995, 1997](#); [Zhao and Takano, 1999](#)) were constrained by the limited computing capability of desktop computers and the lack of large amount of high-quality training data. The majority of the early ANNs had only one or two

Figure 2. Training data templates and components of confusion matrix. (a) An example of input three-component seismograms and its target functions. (b) The criteria and examples of different cases in the confusion matrix. When an arrival residual between the manual arrival and its corresponding predicted arrival is less than 0.1 s, it is considered a true pick. A predicted phase pick with a probability larger than 0.5 is viewed as a positive pick. The color version of this figure is available only in the electronic edition.

hidden layers, and the number of neurons in each hidden layer was often less than 10. The number of seismograms in training datasets varied from a few dozens to a few hundreds. Some of the early ANN-based automatic seismic phase-picking implementations produced large numbers of false picks (i.e., noise was picked as regular seismic phases) and inaccurate arrival times. The overall performance of the early ANNs was not much better than that offered by the conventional algorithms based on characteristic functions, and the anticipated advantages of ANNs were not materialized.

Recent advances in parallel computing technology, in particular, the wide adoption of the graphic processing unit for general-purpose computing ([Chetlur et al., 2014](#)), have substantially increased the computing capabilities of modern commodity computers, thereby, opening up the possibilities of using deep neural networks with tens to hundreds of hidden layers for practical applications ([Hinton et al., 2012](#); [Krizhevsky et al., 2012](#); [Goodfellow et al., 2016](#)). If we treat the time axis of the seismogram as one spatial axis, seismograms can be considered as 1D images and the various deep neural networks designed for 2D image processing, such as the convolutional neural network (CNN) and its variants,

can be adapted to analyze seismograms (Perol *et al.*, 2018; Ross, Meier, and Hauksson, 2018; Ross, Meier, Hauksson, and Heaton, 2018; Kong *et al.*, 2019; Zhu *et al.*, 2019; Zhu and Beroza, 2019). Compared with the early implementations based on shallow ANNs, the overall performance of this new generation of CNN-based deep-learning systems for automatic seismic phase picking is highly encouraging (Zhu *et al.*, 2019; Zhu and Beroza, 2019). In this report, we document our recent progress in adapting the U-Net (Ronneberger *et al.*, 2015), a CNN-based encoder–decoder architecture with long skip connections designed for semantic image segmentation (Chen *et al.*, 2017), to automate the seismic phase-picking task. In particular, we report the performance improvement in the robustness and the accuracy of our autopicking system brought by incorporating the soft attention gate (AG; Schlemper *et al.*, 2019) and the recurrent-residual convolution unit (RRCU; Liang and Hu, 2015) into the baseline U-Net architecture.

Methods

Some of recent deep-learning studies on seismograms (Perol *et al.*, 2018; Ross, Meier, and Hauksson, 2018; Ross, Meier, Hauksson, and Heaton, 2018; Kong *et al.*, 2019; Zhu *et al.*, 2019; Zhu and Beroza, 2019) have mainly focused on the application of the CNN and its variants by treating the seismogram as a 1D image. However, a fundamental difference between the seismogram and an ordinary 2D image is that there are no preferential directions on an ordinary image, whereas, the time axis of the seismogram represents a natural direction of the flow of information, and the different seismic phases appear along the time axis in a specific order that is dictated by the physics of subsurface seismic wave propagation. The similarity between seismograms and sounds (Holtzman *et al.*, 2013; Paté *et al.*, 2016, 2017; Boschi *et al.*, 2017) suggests that the seismogram data are amenable to various machine-learning techniques designed for sequential data (Graves *et al.*, 2006, 2013; Foggia *et al.*, 2015; Chan *et al.*, 2016; Jaitly *et al.*, 2016; Chiu and Raffel, 2017; Adavanne *et al.*, 2018; Mousavi *et al.*, 2020). One of the motivations for our work is to provide an alternative scheme, to account for long-range dependencies among small-scale features on the seismograms for phase picking. In the U-Net architecture, the long-range context is obtained by going to the deeper layers, where units have larger receptive fields, and the small-scale features are preserved through skip connections. However, this approach decouples the two considerations: scale and range, and may not be optimal if long-range context is important for recognizing small-scale features at the same level. The recurrent neural network (RNN) and its variants provide the complementary approach for building long-range context, without compromising small-scale information. By incorporating recurrent units into the U-Net architecture, we can capture the “same-level” context without going into the deeper levels. This approach does not decouple range from scales. The RNN serves a similar

purpose as, but is more general than the autoregressive modeling technique that has already been adapted in some conventional automatic seismic phase-picking algorithms, such as AR-AIC (Sleeman and Van Eck, 1999). Compared with the conventional autoregressive process, the RNN introduces non-linearity into the model and removes the need for specifying the lags for predicting the next value, thereby, offering the potential for modeling sophisticated time-dependent patterns of a variety of seismic waveforms.

Input and output of the neural network

We denote the input seismogram data to our neural network as $x^{(s,r,c)}(t)$, in which the triplet (s, r, c) represents a unique combination of the seismic source index s , the receiver (seismometer) index r , and the component index c (i.e., the ground motion at the receiver r at a given time t is often recorded as a vector with three components: east–west, north–south, up–down). The temporal index t can be represented as integers, if we introduce a fixed time sampling interval Δt , which can be determined based upon the sampling intervals of the seismometers. The input seismogram data can then be represented as a matrix \mathbf{X} , with its columns of a certain row $\mathbf{x}(t)$ representing the recorded ground-motion amplitudes at all components of the receiver(s) (arranged according to a prespecified order) at a given temporal index t and with its rows of a certain column representing the recordings of a component of a certain receiver at all temporal indexes $t \in [t_0, t_0 + T]$, in which t_0 is the beginning temporal index, and T is the duration of the recordings.

The output sequence of our neural network $\mathbf{y}(t)$ is aligned with the input seismogram data along the time interval $[t_0, t_0 + T]$. We denote each element of the output vector $\mathbf{y}(t)$ as $y_D^{(s,r)}(t) = \Pr(t = \tilde{t}_D^{(s,r)} | \mathbf{X}, \mathbf{H})$, which is the conditional probability of temporal index t being equal to $\tilde{t}_D^{(s,r)}$, that is, the true arrival time of the seismic phase $D \in V$ for the source–receiver pair (s, r) , given the input seismogram data \mathbf{X} and the hidden state (i.e., memory) \mathbf{H} of the neural network. The vocabulary V is the set of all seismic phases we are trying to pick. For regional seismic studies, we can choose $V = \{P, S, \text{noise}\}$, in which the “noise” category (Zhu and Beroza, 2019) is equivalent to the null category (Graves *et al.*, 2006) used in speech recognition. A direct consequence of this definition is that for a given source–receiver pair (s, r) , the summation of corresponding elements in the output vector at every temporal index t should be unity, that is, $\sum_D y_D^{(s,r)}(t) = \sum_D \Pr(t = \tilde{t}_D^{(s,r)} | \mathbf{X}, \mathbf{H}) = 1$, which can be implemented by passing the logits of the last layer through the softmax activation function before output. If the arrival times of more seismic phases (e.g., the surface wave) need to be picked, we can expand the vocabulary set V , accordingly. For a given source index s , the number of elements in vector $\mathbf{y}(t)$ is then $N_r \times N_D$, in which N_r is the number of receivers, and N_D is the number of categories in the vocabulary V . The

ordering of the elements in $\mathbf{y}(t)$, with respect to the receiver index r , is identical to that used in the corresponding input seismogram data $\mathbf{x}(t)$. The estimated arrival time of phase D for the source–receiver pair (s, r) can then be obtained from the temporal index that maximizes the corresponding conditional probability, that is, $t_D^{(s,r)} = \operatorname{argmax}_{t \in [t_0, t_0 + T]} y_D^{(s,r)}(t)$.

Multiscale and long-range linkages among seismic phases

For a given source–receiver pair (s, r) , the appearances of different seismic phases along the time axis follow a predictable order that is determined by the physics of subsurface seismic wave propagation. For regions with well-calibrated subsurface Earth structure models, such predictability has been demonstrated to a remarkable level of accuracy (Lee, Chen, and Jordan, 2014; Lee, Chen, Jordan, *et al.*, 2014; Chen and Lee, 2015). In particular, synthesized seismograms obtained using a method-of-lines technique, such as the time-domain finite-difference method for solving the 3D elastodynamic partial-differential equation, can reliably predict the actual observed seismograms, wiggle-for-wiggle, before the actual occurrences of the earthquakes (Lee, Chen, Jordan, *et al.*, 2014). Such predictability of the sequential ordering of the seismic phases exists at multiple scales (i.e., frequencies or periods when considering data on the time axis). For regional seismic studies, an example of large-scale sequential ordering of seismic phases is that the P wave always appears earlier than the S wave, that is, $y_S^{(s,r)}(t) = 0, \forall t \leq \tilde{t}_P^{(s,r)}$. In the reverse time direction, we have $y_P^{(s,r)}(t) = 0, \forall t \geq \tilde{t}_S^{(s,r)}$. Examples of small-scale sequential ordering of seismic phases may involve scattered waves that form the coda waves (Aki and Richards, 2002) following the direct-arriving waves, for example, waves scattered once usually arrive at the receiver earlier, with larger amplitudes, than waves scattered twice, and so on. In our current study, we do not consider the possibility of picking the scattered waves automatically using neural networks. However, we point out that, in a previous study (Lee and Chen, 2013), we were able to automatically pick scattered seismic phases by exploiting the multiscale nature of seismic phases using the continuous wavelet transform and the topological watershed (Vincent and Soille, 1991) image segmentation algorithm.

The temporal separations among different seismic phases depend upon the source–receiver distance, as well as the subsurface Earth structure sampled by the wavepaths of those phases. For a uniform structure, the temporal separation between the P and S waves is $\tilde{t}_S^{(s,r)} - \tilde{t}_P^{(s,r)} = d(1/v_S - 1/v_P)$, in which d is the distance between source s and receiver r , v_P , and v_S are the speeds of the P and S waves, respectively. For regional seismic studies, in which d can reach hundreds of kilometers, the separation between P and S waves can easily exceed tens of seconds for typical wavespeeds in the crust (e.g., $v_P = 6.5$ km/s, $v_S = 3.5$ km/s). Compared with the arrival times, which are abrupt waveform changes with a time scale ~ 0.1 s, the temporal

separation between P and S waves is about two orders-of-magnitude larger. In the framework of the CNN, long-range linkages among different parts of data can be accounted for by increasing the depth of the network, but at the expense of losing small-scale resolution in the deeper layers due to the expansion of the receptive field with network depth and downsampling in the pooling layers. For seismic phase picking, small-scale features are highly important for the accuracy of the estimated arrival times. As an example, under certain scenarios for d larger than ~ 200 km, the first-arriving P wave may become the headwave (P_n), which is usually a small-scale (short-period), low-amplitude (poor SNR) wave propagating along the interface between the crust and the mantle (i.e., the Mohorovičić discontinuity; Aki and Richards, 2002). The temporal separation between the P_n and the S waves can be even larger than that between the ordinary P and S waves for the same source–receiver distance, because a major portion of the P_n propagation wavepath is in the mantle, which usually has a higher v_P value than the crust. It is not uncommon, even for experienced seismologists, to completely miss the P_n phase and misidentify a scattered wave as the first-arriving P wave, when picking the phases manually for such difficult scenarios.

The CNN provides an efficient hierarchical framework for representing multiscale information in a wide variety of data, including the seismogram data. However, it is often cumbersome to use the CNN alone for capturing long-range linkages among small-scale or mixed-scale features, which are important for automatically picking seismic phases that are intimately related with each other through the underlying physics, but have large separations among them along the time axis. In the CNN, the layers responsible for recognizing small-scale features may not be aware of the long-range linkages among them, whereas the layers capable of capturing the low-level features may not have sufficient resolution to detect small-scale features. A straightforward approach for addressing this shortcoming of the CNN is to incorporate the RNN, or its variants, into each convolutional layer of the CNN architecture, to extend its capability of recognizing long-range linkages among features of different scales (Fig. 3a). In this study, we adapt the 2D formulation of the recurrent-residual convolutional network developed for image classification or segmentation (Liang and Hu, 2015; Alom *et al.*, 2018) to the 1D seismogram data, by treating the time axis of the seismogram as a spatial axis. For the i th unit on the k th feature map in a recurrent convolutional layer l , the net input $z_{kl}^{(q+1)}(i)$ for recurrent iteration $q + 1$ is

$$z_{kl}^{(q+1)}(i) = (\mathbf{W}_{kl}^F)^T \mathbf{x}_{l-1}(i) + (\mathbf{W}_{kl}^{(q)})^T \mathbf{h}_{kl}^{(q)}(i) + b_{kl},$$

in which $\mathbf{x}_{l-1}(i)$ is the i -centered patch inside the feedforward input from the previous layer $l - 1$, and $\mathbf{h}_{kl}^{(q)}(i)$ is the i -centered patch inside the recurrent input of the current layer l , \mathbf{W}_{kl}^F and $\mathbf{W}_{kl}^{(q)}$ are the corresponding weights for the feedforward and recurrent connections, and b_{kl} is the bias term. The first term in the equation is identical to the one used in the standard

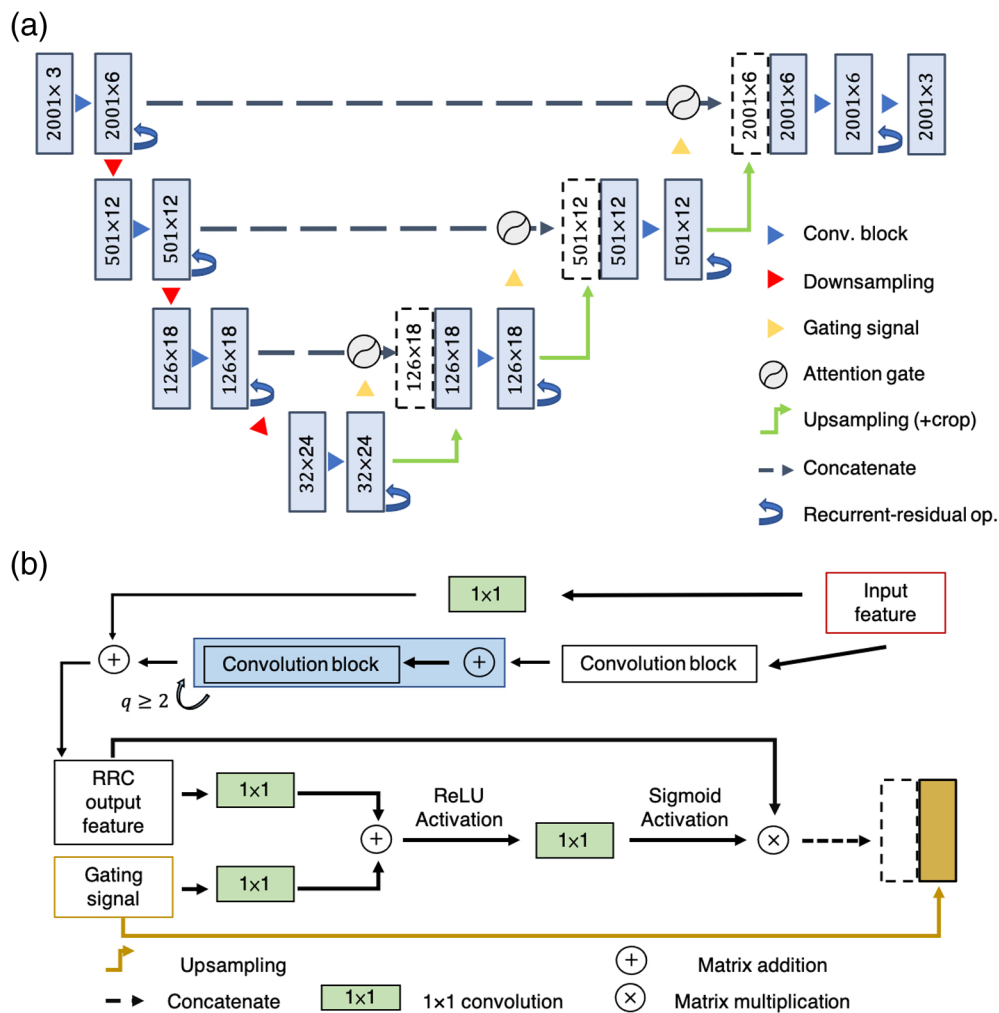


Figure 3. The seismic phase-picking model constructed in this study. (a) Recap of attention recurrent-residual U-Net (attention gate [AG] with recurrent-residual convolution unit [ARRU]-Net) model. The recurrent-residual convolution operation is conducted following every convolution block. AGs re-adjust weightings of feature maps passing through skip connection, which are processed by recurrent-residual modules. The shape of input data and feature maps are shown in rectangles, representing length \times number separately. (b) Schematic diagram of attention recurrent-residual convolution (ARRC) module. In this study, we iterate the recurrent-residual convolution blocks by three times (i.e., $q = 3$ in equation A1). The soft AG module adjusts the weightings of the recurrent-residual convolution unit (RRCU) output feature maps passing through long skip connection in backbone U-Net, with gating signal referring to high-level feature maps from the decoder. The color version of this figure is available only in the electronic edition.

CNN. The second term is induced by the recurrent connections, and the state $\mathbf{h}_{kl}^{(q)}$ evolves with recurrent iteration q (Liang and Hu, 2015). A more detailed formulation for our 1D seismic application is given in the Appendix.

Localization around arrival times of target seismic phases

The accuracy and robustness of the automatic seismic phase-picking algorithm depends upon how to treat the noises that interfere with the onset signals of the seismic phases. The

noises can come from a variety of sources, which may have a wide range of frequencies (i.e., scales). In the actual seismic phase picking, seismologists focus on observing the changes (in amplitude and/or frequency) in seismograms, instead of keeping a consistent focus on entire recordings. The attention mechanism is widely used in deep learning to utilize the input information efficiently. There are two main attention mechanisms. The general attention can assign different weights to input features, to highlight the important parts (e.g., Schlemper et al., 2019), and the self-attention can learn the important relationships between different parts in input data (e.g., Mousavi et al., 2020). To suppress the interference of noises at different scales and improve localization that highlights the onset signals of seismic phases, we adapted in our neural network the grid-based soft-attention gating module (Schlemper et al., 2019) that had been integrated into the CNN and its variants, such as, the U-Net, for image classification and segmentation applications. However, a fundamental difference in our adaptation is that both the input features and the gating signals, which is gathered from a coarser level used by the AGs in our neural network (Fig. 3b) are generated

from the recurrent-residual convolutional layers, rather than the standard convolutional layers. In particular, the k th channel in the feature map $\mathbf{x}_{l+1}(i)$ can be written as $x_{kl}(i) + f[z_{kl}^{(Q)}(i)]$, in which $f[\cdot]$ is the activation function, and $z_{kl}^{(Q)}(i)$ has contributions both from the feedforward connections and from the recurrent connections when $Q > 1$ (equation A1). To separate the contribution of the recurrent connections to the calculation of the attention coefficients, we split the feature map into two components, that is, $\mathbf{x}_{l+1}(i) = \mathbf{x}_{l+1}^{(0)}(i) + \mathbf{x}_{l+1}^{(Q)}(i)$, in which $x_{kl}^{(0)}(i) = x_{kl}(i) + f[z_{kl}^{(1)}(i)]$ is the

feature map without the recurrent connections and $x_{k(l+1)}^{(Q)}(i) = f[z_{kl}^{(Q)}(i)] - f[z_{kl}^{(1)}(i)]$ is the contribution to the feature map made by the recurrent connections, and, we have $x_{l+1}^{(Q)}(i) = 0$, when $Q = 1$.

One of the main difficulties in fully automating the seismic phase-picking task using conventional algorithms is in separating the interference of various seismic noises (e.g., ocean waves, human activities, winds), which are usually multiscale and nonstationary. An example is the STA/LTA algorithm based on the ratio of the short-time and long-time moving averages for detecting abrupt changes on the seismogram. In practice, the widths of the short-time and long-time moving windows need to be carefully selected, based on the common spectral characteristics of the seismogram data under study. In particular, the width of the short-time window needs to be wide enough to smooth out high-frequency noises (e.g., anthropogenic activities, wind), whereas the width of the long-time window needs to be narrow enough to reject long-period noises (e.g., tidal waves, teleseismic waves). Many of the preprocessing procedures used in conventional algorithms are designed to suppress multiscale seismic noises and often require the most amount of human effort in parameter turning. Some of the studies have been focusing on adapting conventional time-domain algorithms for phase picking to the time–frequency domain (Zhang *et al.*, 2003; Galiana-Merino *et al.*, 2008; Lee and Chen, 2013; Lee *et al.*, 2019), to remove some of the preprocessing steps aimed at suppressing multiscale noises. The grid-based soft AG (equations A2 and A3) coupled with the U-Net encoder–decoder architecture (i.e., the attention U-Net; Schlemper *et al.*, 2019) provides a natural, automatic, trainable, and efficient mechanism, for focusing on the onset features of target seismic phases while suppressing the interference from noises at each scale. In particular, the grid-based gating signal at the coarse scale encodes the context in a temporal neighborhood surrounding the onset signal, and the additive attention (equation A3) enhances onset features that are consistent with the gating signal across scale, whereas noises that tend to be incoherent across scale are being suppressed. The attention-gated feature map x_l contains only activations that are relevant to the onset signal and is then concatenated into the decoder through the skip connection (Fig. 3a).

Data

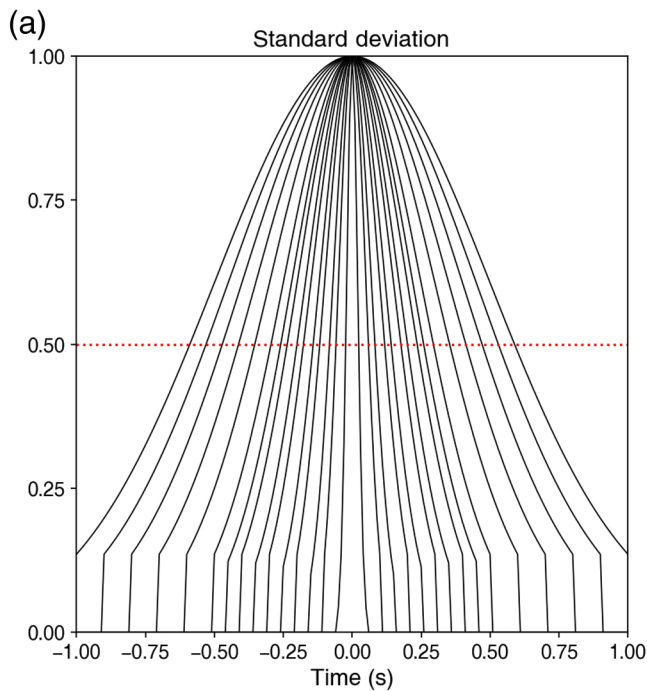
In this study, seismograms of earthquakes that occurred around Taiwan and recorded by Central Weather Bureau (CWB) network and Broadband Array in Taiwan for Seismology are used as the waveform dataset for our machine-learning algorithm. Taiwan is located at an ongoing arc–continent collision zone, where the Philippine Sea plate (PSP) converges to the Eurasian plate (EP) with a rate of about 56–82 mm/yr (Yu *et al.*, 1997). In the north end, the PSP subducts northward below the EP along the Ryukyu trench, and the back-arc rifting basin, Okinawa trough, has also extended

into northern Taiwan. In the south end, the oceanic South China Sea plate subducts eastward below the PSP along the Manila trench. High seismicity reflects the complex tectonic settings near Taiwan (Fig. 1a,b). The abundance of earthquake recordings with diverse focal mechanisms, wide ranges of magnitudes as well as hypocenter depths, and relatively noisy recording environments in Taiwan (e.g., anthropogenic activities and natural ambient noises such as tides, surface runoffs, and so forth) enable to examine the reliability and robustness of our phase-picking algorithm. We gathered and chronologically split 1,052,675 sets of three-component seismograms from year 2012 to 2019 (Fig. 1); seismograms gathered from 2012 to 2017 are used for model training (603,054; about 57.29%) and validation (150,764; about 14.32%), the rest from 2018 to 2019 are utilized for model testing (298,857; about 28.39%).

In this study, the input three-component seismograms are 20 s long, with one set of *P*- and *S*-phase labeling (Fig. 2a). To ensure *P* and *S* waves' features could be captured by the algorithm, we keep, at least, 3-second-long waveforms before labeled *P* arrivals and 5-second-long waveforms following labeled *S* arrivals. The position of the phase arrivals is randomly distributed within a 20 s waveform, to prevent the influence of the windowing during training. The longest allowable *P*- and *S*-arrival differential time is 12 s, and the length of time is sufficient for most recordings of regional earthquakes (Fig. 1). Under the scheme of training a deep neural network, quality and consistency of inputs (*P* and *S* picks) play essential roles in model performance. The *P* and *S* arrivals used in this study are manually picked by well-trained experts in CWB, and the picks are divided into four levels according to their picking qualities. In this work, we only selected the seismograms with the top qualities of *P* and *S* picks as data for method verifications. We apply *Z*-score standardization to the seismograms to reduce the wide variances in the amplitude. Each input waveform first removes its mean and then is divided by the standard deviation. Data standardization restricts the range of data values by their means and standard deviations, and ensures that the dataset is internally consistent. In machine learning, data standardization could make input data have similar contributions during training and could speedup model convergence.

Target function prototypes

To adapt the U-Net architecture to address the automatic seismic phase-picking task, we design our neural network to map the input seismogram to a set of probabilistic target functions (Graves *et al.*, 2006). Each target function is a time series with the same dimension as the input seismogram and indicates the probability of occurrence of a particular seismic phase as a function of time (Graves *et al.*, 2006). The maximum of the target function is the arrival time of the corresponding seismic phase. We classify the probability at each time sample on the seismogram into three categories: a *P*-wave arrival, an *S*-wave



arrival, and the other waves. The probability summation of the three target functions at each time sample must be one, which can be implemented using the softmax activation function in the output layer of our neural network.

The ground-truth target functions of the *P*- and *S*-wave arrivals used in training, validation, and test datasets are constructed by the manual arrival picks provided by CWB. In this study, the target function is a zero-padded truncated Gaussian distribution, with its peak located at the arrival time of the seismic phase (Fig. 2a). The loss function used for training can then be defined in terms of the cross entropy between the softmax normalized predicted target functions and the corresponding ground-truth target functions (Zhu and Beroza, 2019). The confusion matrix is adapted to compare the performances among models. In our confusion matrix, both the absolute time residual between a manual pick and its corresponding predicted arrival, $\tilde{R} = |t_D^{(s,r)} - \hat{t}_D^{(s,r)}|$, and the probability of a phase at its predicted arrival, $\hat{y} = y_D^{(s,r)}(t_D^{(s,r)})$, are considered. When a pick with $\tilde{R} < 0.1$ s, it is regarded as a true pick; when a pick with $\hat{y} > 0.5$, it is regarded as a positive pick.

The width (i.e., standard deviation) of the truncated Gaussian distribution is closely related to the label-smoothing regularization technique for tackling overfitting and overconfidence (Goodfellow et al., 2016; Szegedy et al., 2016). If the Gaussian is too narrow, the model will be less adaptive and overly confident about its predictions, which can lead to overfitting. In principle, the width of the Gaussian functions for different seismic phases should be determined from the uncertainties of their manually picked arrivals. However, in practice, we usually do not have independent estimates of such uncertainties. To properly estimate the width of Gaussian functions

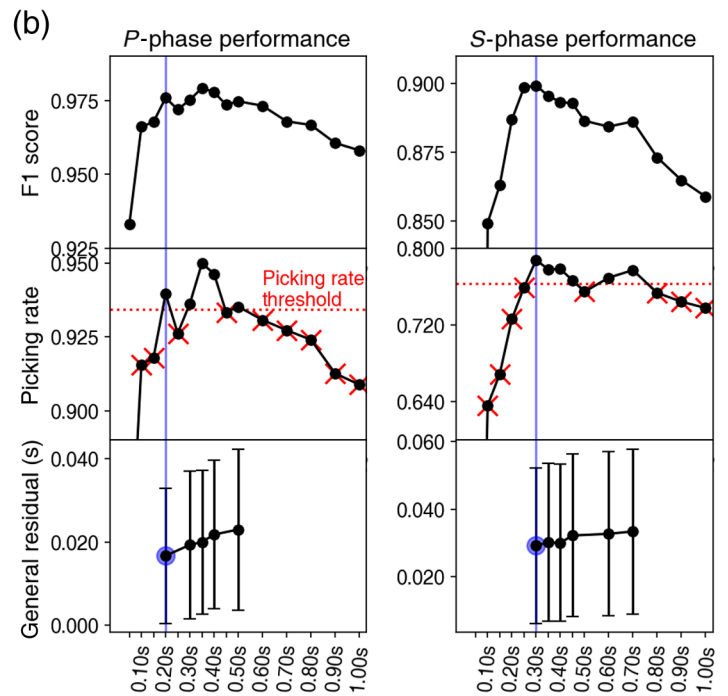
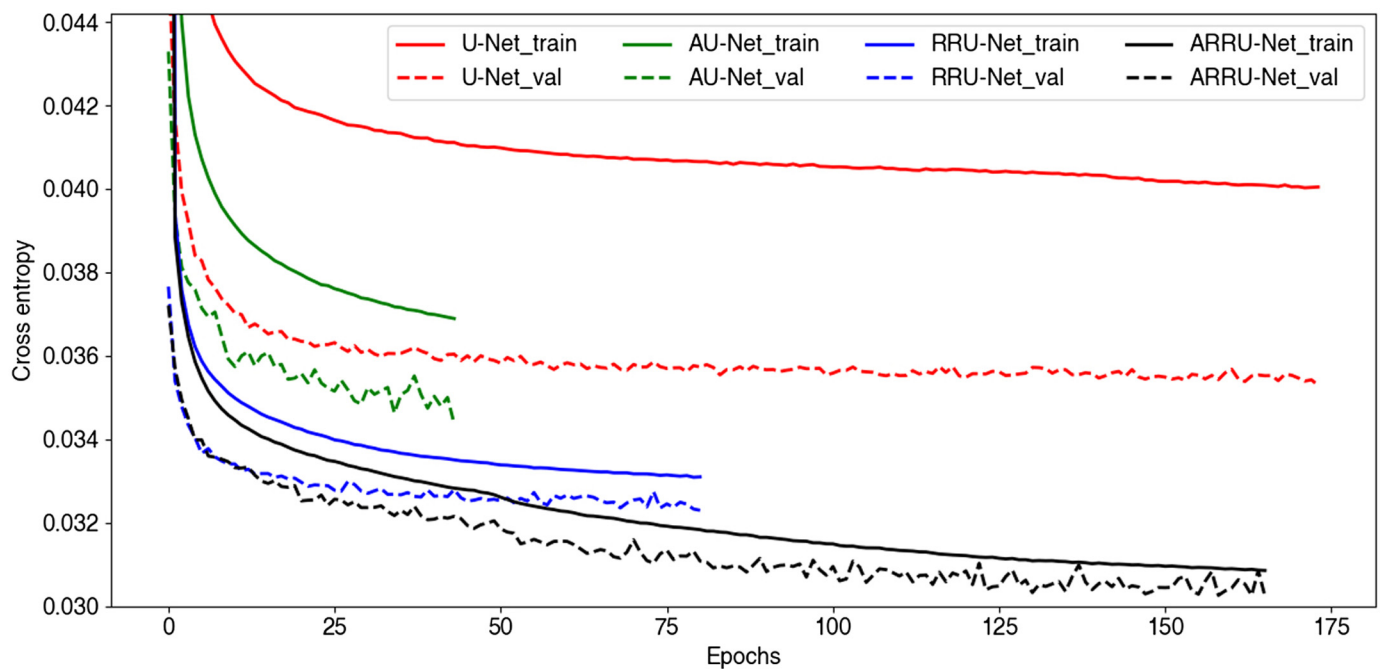


Figure 4. Decision of truncated Gaussian distribution width centering at target seismic phases. (a) Truncated Gaussian functions with standard deviations ranging from 0.05 to 1.0 s used for *P* and *S* target function tests in this study. (b) The F1 scores (top), picking rates (middle), and mean arrival residuals (bottom) of *P* (left) and *S* (right) arrivals on tested target functions. The picking rate threshold is the median of all testing results. The one with the smallest mean arrival residual among the qualified candidates is selected as the target function of the phase. The color version of this figure is available only in the electronic edition.

for a seismic phase, we adopt a model calibration approach that takes into account the prediction performance metrics, including the picking rate and the arrival residual R . The picking rate is defined as the rate of true-positive picks, and arrival residual refers to the time residual of true-positives picks universally categorized by all models under comparison. We built 15 sets of models using U-Net, with width of the truncated Gaussian distribution from 0.05 to 1.0 s for both *P* and *S* target functions (Fig. 4). Let O be the set of models constructed, J_D^o as picking rate of model $o \in O$, and \bar{J}_D^o the mean value of all models' picking rate on seismic phase D . Picking rate threshold K_D is defined as: $K_D = \text{MEDIAN}(\{J_D^o | \forall J_D^o \geq \bar{J}_D^o\})$. Optimal width L_D is decided by general residual t_D^o , which can be expressed as: $L_D = \text{argmin}(\{t_D^o | \forall J_D^o \geq K_D\})$. The calibrations show that the optimal standard deviation of Gaussian functions is 0.2 s for the *P* picks and 0.3 s for the *S* picks in our dataset (Fig. 2a). The calibration results are consistent with our common sense that *S* picks usually have larger uncertainties than that of *P* picks due to the perturbations of *P* coda or converted waves during picking *S* arrivals.



Results

Training and learning curves

In this study, the Rectified Adam optimizer was used with a constant learning rate (1×10^{-4}) for optimizing the models, and we did not implement model regularization nor data augmentation techniques. In the model training, we set the dropout rate to 0.1 for all dropout layers, and the early stopping criterion was that the validation loss did not improve for 20 successive epochs. The training times per each epoch with 603,054 training and 150,764 validation data using single NVIDIA RTX 2080Ti were 1034 s for the standard U-Net, 3526 s for the U-Net with AGs (AU), 2640 s for the U-Net with RRCU (RRU), and 5006 s for U-Net with both AGs and RRCU (ARRU). Figure 5 shows the train and validation learning curves of a standard U-Net, AU, RRU, and ARR-Net models over training epochs of the seismic phase-picking task. In this study, we stop the model training when the validation loss is no longer improving, and the U-Net stops at the 174th epoch, AU stops at the 44th epoch, RRU stops at the 81th epoch, and ARR-Net stops at the 166th epoch. Generally speaking, a model should be first checked whether it is overfitting, which indicates that a model fits training data too well but cannot generalize to unseen data, validation dataset. Though there is no strict definition, the intuitive phenomenon for overfitting turns to be condition that training loss keep decreasing with increasing validation loss. The difference between training and validation loss implies the performance gap of a model between training and validation datasets. In general, a good-fit model is identified by stably decreasing, both in training and validation loss, to a convergence point with a minimum gap. Among the models, the ARR-Net has the lowest loss value, and the minimum loss gap between its train and validation learning curves, showing the outstanding capability of fitting training data.

Figure 5. Learning curves of U-Net, attention U-Net (AU-Net), recurrent-residual U-Net (RRU-Net), and ARR-Net built in this study. Solid lines represent training loss, and dashed lines indicate validation loss. The color version of this figure is available only in the electronic edition.

Model performances

Table 1 lists the seismic phase-picking results of machine-learning algorithms tested in this study, including the U-Net, AU, RRU, and ARR-Net models and the AR-AIC picker, which is a conventional automatic phase-picking method combining STA/LTA and AIC for seismic phase picking, implemented by ObsPy (Sleeman and Van Eck, 1999; Krischer *et al.*, 2015). The precision, recall, and F1 score are computed for evaluating machine-learning-based models. The picking rate is defined as the number of phase picks detected by the algorithm (true-positive picks for the machine-learning-based models), divided by the total number of manual picks in the test dataset. The means and standard deviations of arrival-time residuals between the detected and manual picks for models are also calculated for model evaluations. Overall, all the methods show that the performance on picking *P* arrivals is better than that on picking *S* arrivals. In this work, we use the same criteria (i.e., phase residual < 0.1 s) for both *P*- and *S*-phase-picking evaluations. However, the *S* picks usually have higher uncertainties due to perturbations (e.g., *P* coda waves) during picking *S* arrivals in practical phase-picking tasks. This phenomenon was also found in the tests for the width of *P* and *S* target functions (i.e., the optimal target function of *S* is wider than that of *P*). It is likely that the uncertainties in *S* arrivals have been internalized in the machine-learning-based models during model training.

TABLE 1

Model Performances on Taiwan Test Dataset (298,857 Sets)

		Mean (s)	St. Dev. (s)	Precision	Recall	F1 Score	MAE (s)	MAPE (s)	Picking Rate
<i>P</i> phase	AR-AIC	−0.0952	0.1840	—	—	—	0.1318	0.0245	0.7695
	U-Net	0.0027	0.0248	0.9749	0.9659	0.9704	0.0180	0.0032	0.9251
	AU-Net	0.0068	0.0239	0.9755	0.9865	0.9810	0.0179	0.0032	0.9519
	RRU-Net	0.0073	0.0219	0.9803	0.9952	0.9877	0.0162	0.0029	0.9714
	ARRU-Net	0.0043	0.0203	0.9797	0.9929	0.9862	0.0129	0.0023	0.9672
<i>S</i> phase	AR-AIC	−0.2351	0.4849	—	—	—	0.4429	0.0420	0.3242
	U-Net	0.0024	0.0436	0.8296	0.9656	0.8925	0.0346	0.0033	0.7675
	AU-Net	0.0008	0.0420	0.8449	0.9707	0.9034	0.0325	0.0030	0.7896
	RRU-Net	0.0029	0.0411	0.8532	0.9858	0.9147	0.0322	0.0030	0.8232
	ARRU-Net	0.0107	0.0350	0.9131	0.9934	0.9516	0.0280	0.0026	0.9007

Mean and standard deviation (st. dev.) of the differences of manual picked arrivals minus model predicted arrivals in seconds. MAE and MAPE are mean absolute difference and mean absolute percent difference, respectively. Bold values represent the best performance. AR-AIC, autoregressive Akaike information criterion; ARRU, attention gate with recurrent-residual convolution unit; AU-Net, attention U-Net; RRU-Net, recurrent-residual U-Net.

Among the U-Net-based implementations, the models with AGs and/or RRCU modules have better performances in picking both *P* and *S* arrivals than that of the U-Net model. These improvements are not only shown on precision, recall, and F1 score, but also reflect on the picking accuracy (i.e., lower means and standard deviations of arrival residuals).

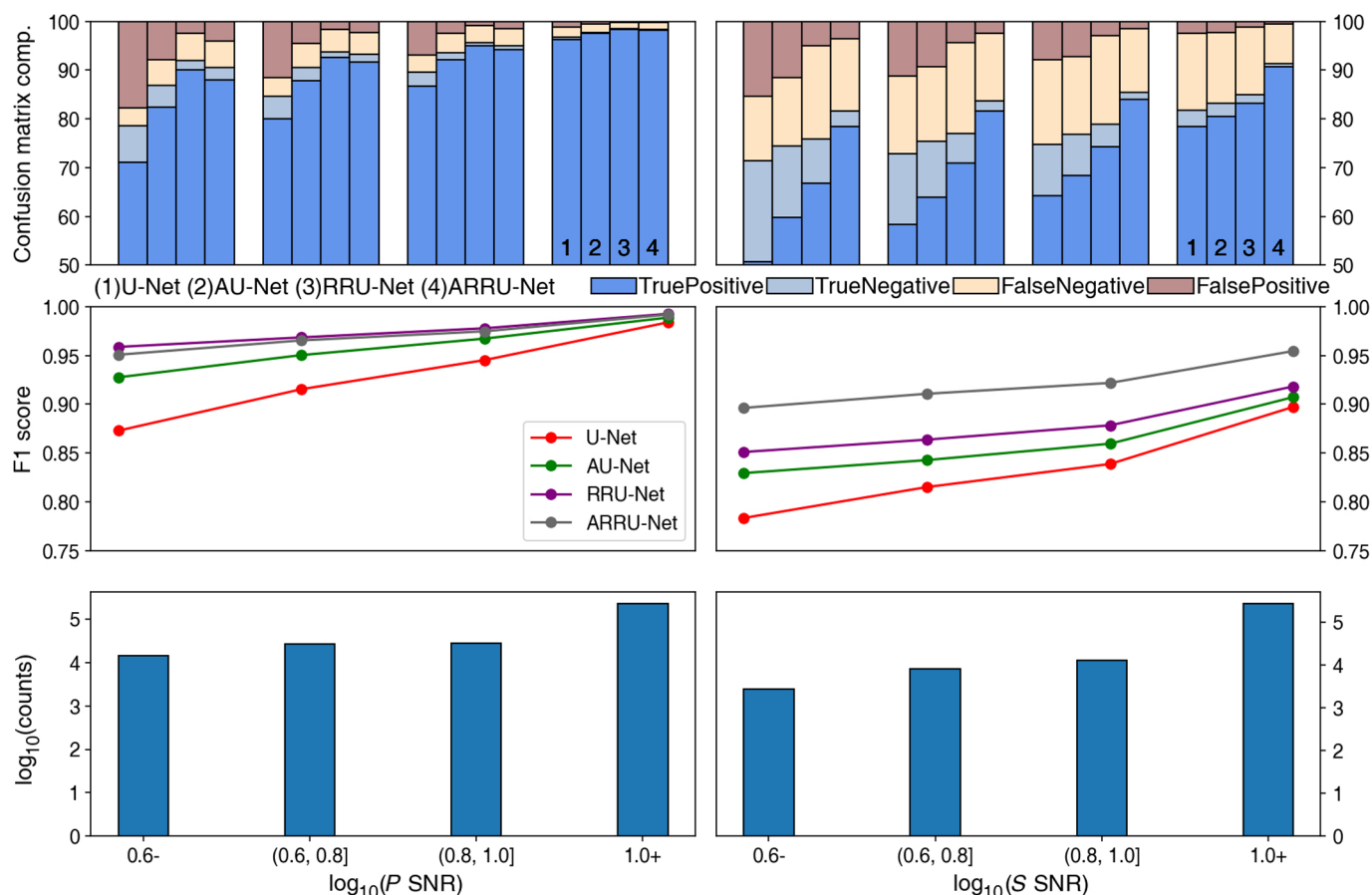
To further explore the influence of the AGs and/or RRCU functions on the seismic picking task, we applied the methods to test dataset grouped by the SNR of the *P* and *S* waves. In this study, the SNR is based on the root mean square of 3 component seismograms. For the SNRs, the noise is a 10 s waveform end at 0.5 s before its manual *P* arrival, the signal of *P* phase is a 3 s waveform start at 0.5 s before its manual *P* arrival, and the signal of *S* phase is a 4 s waveform start at 0.5 s before its manual *S* arrival. Figure 6 shows the performances of the model in different SNR conditions. As expected, the results show that as SNR increase, the models perform better on picking both *P*- and *S*-phase arrivals. Under similar SNRs, all models perform better on picking *P* arrivals than on picking *S* arrivals (Fig. 6). Comparing the performances between the U-Net and AU in picking *P* and *S* arrivals, the improvements are more significant when the SNRs of the phases are relatively low (Fig. 6). The improvements mainly come from the increase of true-positive picks, and the decrease of true-negative and false-positive picks (Fig. 6). When the U-Net was incorporated with the RRCU function, further increases in the F1 score for both picking *P* and *S* arrivals at different SNR ranges could be found (Fig. 6). Probably, the RRCU function can incorporate dynamic changes on a time series (i.e., seismograms), and, therefore, the features on temporal waveform changes could be “learned” for phase-picking tasks, especially on *S* arrivals. When the AGs are added to the RRU, the performances in

picking *P* arrivals are about the same. However, comparing the performances of the RRU and ARRU in picking *S* arrivals, the ARRU shows notable improvements at different SNR ranges. Since picking *S* arrivals from seismograms are more complicated than picking *P* arrivals, probably the combination of AGs and RRCU can better demonstrate their performance, when the data are complicated.

In model comparisons, the ARRU always performs better than the U-Net, especially for picking seismic phases at low SNRs (Fig. 6). Figure 7a–c shows some waveform examples that *P* and/or *S* arrivals can be predicted by ARUU but the U-Net. In the examples, the amplitude changes at *P* and/or *S* arrivals are relatively nuclear, and, therefore, the incorporated AGs and RRCU demonstrate their effects. To some extent, routine phase picking is a laborious and monotonous job, and the quality of manual picking may be affected by different factors, such as accumulated experience, subjective judgments, professionalism, work status, and so on. In addition to providing more reliable predictions of *P* and *S* arrivals at low SNRs, the ARRU can also provide more stable quality on phase picking. Figure 7d–f shows the obvious mistakes of manual phase picking, and the ARRU can provide more reasonable phase picks.

Model generalization

In machine learning, the model generalization, a trained model has about the same performance for unseen data, is one of the most important considerations in practical use. In this work, the trained models are validated using the validation dataset, a set of unseen seismograms recorded in the same time period of training data. Because the seismograms in the validation and training datasets have the same recording period, we, therefore, randomly selected about the same amount of seismograms in



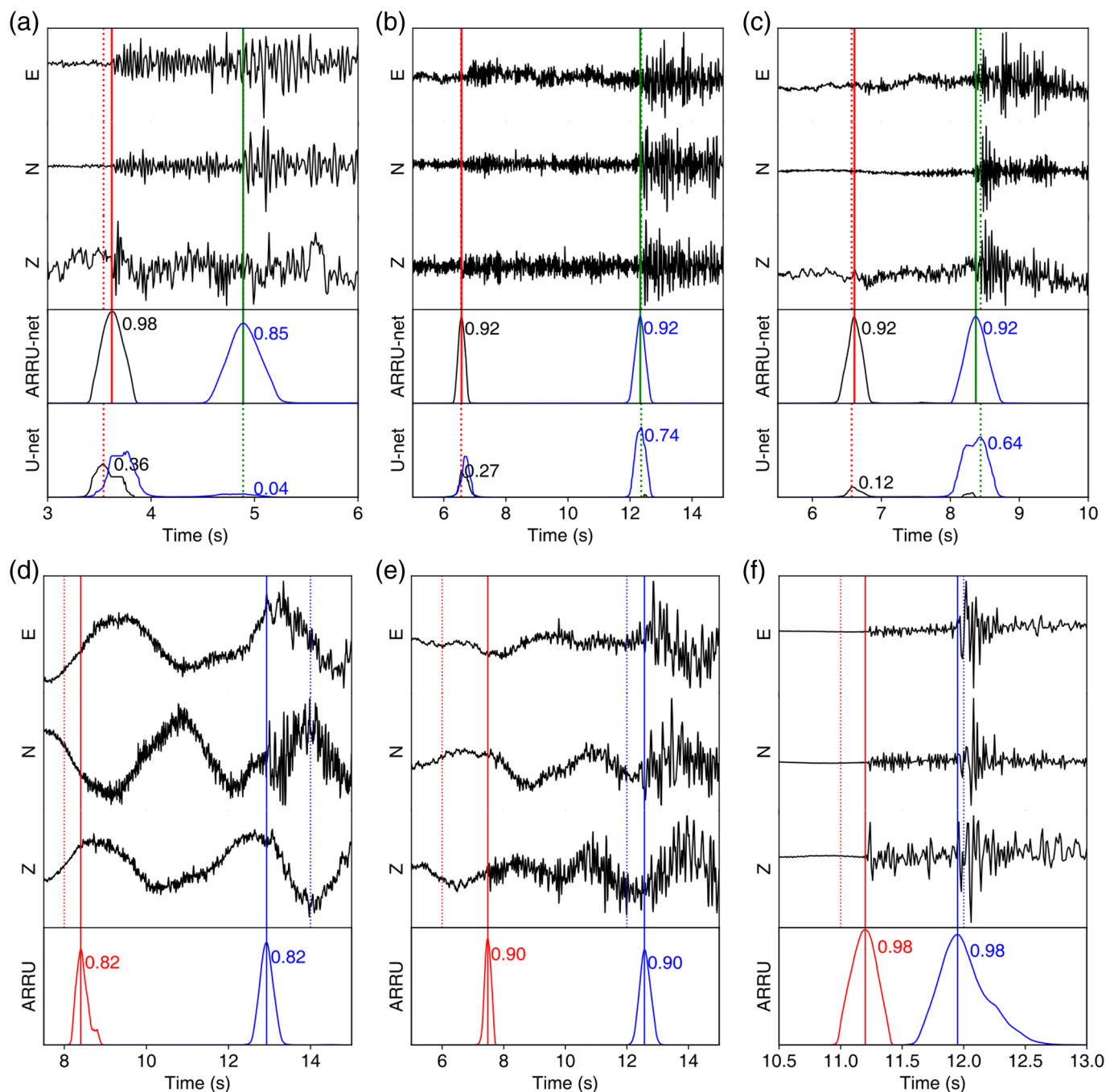
the test dataset, which have no time overlap with training data, for a further model generalization test. The differences of F1 score between the two datasets for models on *P*- and *S*-phase-picking tasks show the generalization capabilities of models (Fig. 8). In general, the F1 score differences on the *P*-phase picking are usually smaller than that of the *S*-phase picking, which means the *P*-phase picking has a better generalization. Among the models, the ARR-Net has the smallest F1 score differences on both *P*- and *S*-wave picking, and reveals the great model generalization of ARR-Net.

To test the generalization capabilities on a distinct dataset, we applied the trained model to earthquake recordings in Southern California (SC). We randomly select 153,055 recordings of earthquakes archived in Southern California Earthquake Data Center catalog (SCEDC, 2013) that occurred between 2010 and 2019 as a dataset for testing our trained models (Fig. 1c,d). The seismograms in SC dataset are processed in the same way as waveforms in the Taiwan dataset. Table 2 shows the performances of the AR-AIC picker, U-Net, and ARR-Net on the SC test dataset. Even if the U-Net and ARR-Net models are trained by only using the seismograms in Taiwan data, the models still can achieve remarkable performances on SC dataset. In fact, the performances of models in SC dataset are better than that in the test dataset of Taiwan, especially the ARR-Net model. The tests show that the machine-

Figure 6. Performance comparisons of different models in picking *P* (left) and *S* (right) arrivals for data in the test dataset with different SNRs. The top shows the confusion matrix components, and the middle shows the F1 scores of U-Net, AU-Net, RRU-Net, and ARR-Net models under different scenarios. The bottom shows the distributions of data in different SNRs. Including AGs and/or recurrent-residual convolution (RRC) functions in U-Net can improve model performances, especially for data with relatively low SNRs. The ARR-Net model shows significant performance improvements in picking *S* arrivals at different SNR levels. The color version of this figure is available only in the electronic edition.

learning-based models, especially the ARR-Net, are able to learn the principles of picking *P* and *S* arrivals from seismograms, and the excellent generalization capabilities of models give us confidences in practical seismic phase-picking tasks.

In addition to the SC dataset, we also tested the ARR-Net phase picker to the Stanford Earthquake Dataset (STEAD)—a global dataset of labeled *P* and *S* arrivals for earthquake recordings (Mousavi et al., 2019). Our following tests are carried out according to the dataset classification in STEAD. We used the same criterion of a true-positive pick (predicted and ground-truth absolute arrival differences less than 0.5 s and the probability a phase larger than 0.3) in Mousavi et al. (2020) in



the following comparisons. To examine the model generalization, we applied the ARR-Net model trained only by the Taiwan waveform dataset to about 90% of waveforms that satisfied our model framework ($|t_S - t_P| < 12$ s) in STEAD test dataset. The ARR-Net model obtained similar performances on both P and S picking tasks as the performances of EQTransformer, which is trained by the STEAD dataset (Table 3). To investigate the performance of our algorithm on another independent dataset, we train and test another ARR-Net model using waveforms in STEAD. The new model uses 1-second-long waveforms before P and 3-second-long waveforms after S arrivals

Figure 7. Examples of ARR-Net predictions compared with U-Net predictions and manual picks. (a–c) Comparison of P (red curves) and S (blue curves) probability functions made using the ARR-Net and U-Net. Solid and dashed lines indicate the qualified phase arrivals predicted by the ARR-Net and U-Net. (d–f) Examples of some problematic manual picks. The comparisons of manual picks (dashed lines) and arrivals predicted by the ARR-Net (solid lines), based on the P (red curves) and S (blue curves) probability functions. The color version of this figure is available only in the electronic edition.

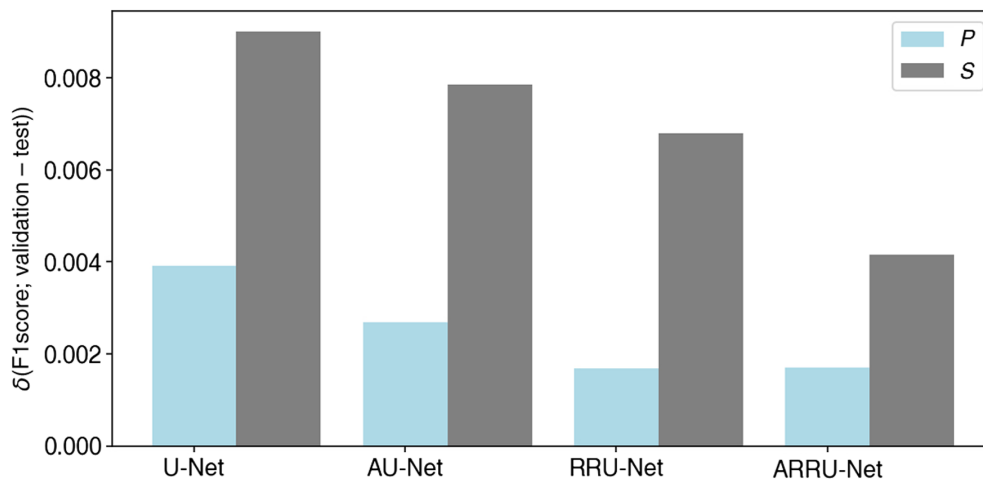


Figure 8. Comparisons of model generalization capability for unseen seismograms in the test dataset. The F1 score differences between the validation and test datasets for *P* (blue bars) and *S* (gray bars) indicate that the ARRU model has a better model generalization. The color version of this figure is available only in the electronic edition.

($|t_S - t_P| < 16$ s) during training and testing, to include more waveforms in STEAD. Data with the length before *P* arrival less than 1 s in STEAD are also included by slicing 20-second-long waveforms from the beginning. The new model uses 90% of the STEAD training data for training and 95% of the STEAD test data for testing. The ARRU model trained by STEAD waveforms used the same training parameters for the Taiwan model, and Table 3 shows the performances of the model. Noted that we did not use the data augmentation, which can potentially improve the model performance (Zhu *et al.*, 2020), used by EQTransformer during training. The ARRU model trained with STEAD data mainly reduces the means and standard deviations of the differences between ground truth and predicted arrivals in STEAD. The performances of the EQTransformer and ARRU phase picker were roughly performed at a similar level. The experiments proved that

different deep-learning architectures could achieve a similar performance level in seismic phase picking.

Discussions

To better understand what the model has learned from data, we applied techniques of Gradient-weighted Class Activation Map (Grad-CAM; Selvaraju *et al.*, 2017), to extract information from the last convolutional layer before softmax activation. The Grad-CAM portraits show the mapped weightings toward the final decision for designated class. Figure 9 shows Grad-CAMs of ARRU and U-Net for the same three-com-

ponent seismograms. On deciding *P* arrival, U-Net model tends to misidentify *S* arrival as *P* arrival, whereas, ARRU is quite confident on making decision around the real *P* arrival. On deciding *S* arrival, ARRU pays most attention around real *S* arrival. U-Net model gives the lowest weightings around *P* arrival, but not as confidential as ARRU does on *S* arrival. This example demonstrates that ARRU is more capable of discriminating features on deciding *P* and *S* arrivals comprehensively than U-Net does.

In practical earthquake locations or tomographic inversions, the consistency of seismic phase pickings will affect the accuracy of the results. Here, we select some earthquakes with high-waveform similarities, and compare the manual picks and phase arrivals made by ARRU (Fig. 10). In comparison, we can find that both *P* and *S* arrivals picked by ARRU show higher consistency. The manual picks are likely affected

TABLE 2
Model Performances on Southern California Dataset (153,065 Sets)

		Mean (s)	St. Dev. (s)	Precision	Recall	F1 Score	MAE (s)	MAPE (s)	Picking Rate
<i>P</i> phase	AR-AIC	−0.0990	0.1403	—	—	—	0.1095	0.0213	0.8139
	U-Net	0.0012	0.0210	0.9850	0.9822	0.9836	0.0148	0.0028	0.9604
	ARRU-Net	0.0025	0.0160	0.9895	0.9954	0.9924	0.0093	0.0018	0.9829
<i>S</i> phase	AR-AIC	−0.2866	0.4296	—	—	—	0.4221	0.0376	0.3311
	U-Net	0.0010	0.0439	0.8520	0.9650	0.9050	0.0348	0.0031	0.7963
	ARRU-Net	0.0109	0.0325	0.9538	0.9963	0.9746	0.0261	0.0023	0.9473

Mean and standard deviation (St. Dev.) of the differences of manual picked arrivals minus model predicted arrivals in seconds. MAE and MAPE are mean absolute difference and mean absolute percent difference, respectively. Bold values represent the best performance. AR-AIC, autoregressive Akaike information criterion; ARRU, attention gate with recurrent-residual convolution unit; AU-Net, attention U-Net.

TABLE 3

Model Performances on STanford EArthquake Dataset (STEAD) (True Pick Threshold = 0.5 s, Positive Pick Threshold = 0.3)

	Model	Mean (s)	St. Dev. (s)	Pr	Re	F1	MAE	MAPE	Training Data	Testing Data	Remarks
P phase	ARRU (TW)	−0.03	0.07	0.99	1.00	0.99	0.04	0.01	Taiwan, 603K	STEAD, 112K	$ t_S - t_P < 12$ s
	ARRU (STEAD)	0.00	0.06	1.00	0.99	0.99	0.03	0.01	STEAD, 1.08M	STEAD, 120K	$ t_S - t_P < 16$ s
	EQTransformer	0.00	0.03	0.99	0.99	0.99	0.01	0.00	STEAD, 1.2M	STEAD, 126K	Mousavi et al. (2020)
	PhaseNet	−0.02	0.08	0.96	0.96	0.96	0.07	0.01	North California, 780K		
	GPD	0.03	0.10	0.81	0.80	0.81	0.08	0.01	South California, 4.5M		
	PickNet	0.00	0.09	0.81	0.49	0.61	0.07	0.02	Japan, 740K		
	PpkNet	−0.01	0.15	0.90	0.90	0.90	0.10	1.90	Japan, 30K		
	Yews	0.07	0.13	0.54	0.72	0.61	0.09	0.02	China, 1.4M		
	Kurtosis	−0.03	0.09	0.94	0.79	0.86	0.08	0.01	—		
	FilterPicker	−0.01	0.08	0.95	0.82	0.88	0.14	0.02	—		
	AIC	−0.04	0.09	0.92	0.83	0.87	0.09	0.01	—		
S phase	ARRU (TW)	−0.04	0.11	0.98	0.99	0.99	0.08	0.01	Taiwan, 603K	STEAD, 112K	$ t_S - t_P < 12$ s
	ARRU (STEAD)	−0.01	0.10	0.98	0.98	0.98	0.06	0.01	STEAD, 1.08M	STEAD, 120K	$ t_S - t_P < 16$ s
	EQTransformer	0.00	0.11	0.99	0.96	0.98	0.01	0.00	STEAD, 1.2M	STEAD, 126K	Mousavi et al. (2020)
	PhaseNet	−0.02	0.11	0.96	0.93	0.94	0.09	0.01	North California, 780K		
	GPD	0.03	0.14	0.81	0.83	0.82	0.10	0.01	South California, 4.5M		
	PickNet	0.08	0.17	0.75	0.75	0.75	0.10	0.03	Japan, 740K		
	PpkNet	0.02	0.15	1.00	0.91	0.95	0.10	1.85	Japan, 30K		
	Yews	−0.02	0.13	0.83	0.55	0.66	0.11	0.01	China, 1.4M		
	Kurtosis	−0.10	0.13	0.89	0.39	0.55	0.11	0.01	—		
	FilterPicker	−0.05	0.13	0.61	0.41	0.49	0.10	0.01	—		
	AIC	−0.07	0.15	0.87	0.51	0.64	0.12	0.02	—		

Mean and st. dev. (standard deviation) of the differences of manual picked arrivals minus model predicted arrivals in seconds. MAE and MAPE are mean absolute difference and mean absolute percent difference, respectively. Bold values represent the best performance. ARRU, Attention Recurrent-Residual U-Net presented in this study; GPD, generalized phase detection; AIC, Akaike information criterion.

by background noise and subjective judgments, and similar cases have also been reported in other data centers (e.g., SC; Shearer, 1997). For a group of waveforms with high similarities, the waveform cross correlations could be used to find accurate relative arrivals for earthquake relocations and/or tomographic inversions (Shearer, 1997; Waldhauser and Ellsworth, 2003; Zhang and Thurber, 2003). However, in most cases, the *P* and *S* arrivals are directly used for inversions. If the criteria for seismic phase picking are inconsistent, the picking errors introduced by subjective judgments will affect the inversion results. In general, the seismic phase arrivals picked by ARRU show better consistency and may reduce the human errors in inversions.

In this study, we made a full comparison of U-Net models on seismic phase-picking tasks and found that additional

modules such as AGs and recurrent-residual blocks could provide performance gain. The model trained by the Taiwan dataset has been applied to various unseen datasets, including the Taiwan test dataset, SC dataset, and STEAD global dataset (Mousavi et al., 2019). The performances of the model are not much different (Tables 1–3), showing the good generalization of the model. In addition, benchmarking is commonly used in machine learning to compare and evaluate machine-learning algorithms. The datasets in STEAD are utilized as an independent benchmark test for our algorithm. Table 3 shows the performance of our ARRU phase picker, and the performances of our algorithm on different datasets are quite consistent. In machine learning, various techniques have been used to improve the performance of an algorithm, such as data augmentation and hyperparameter optimization (Li et al.,

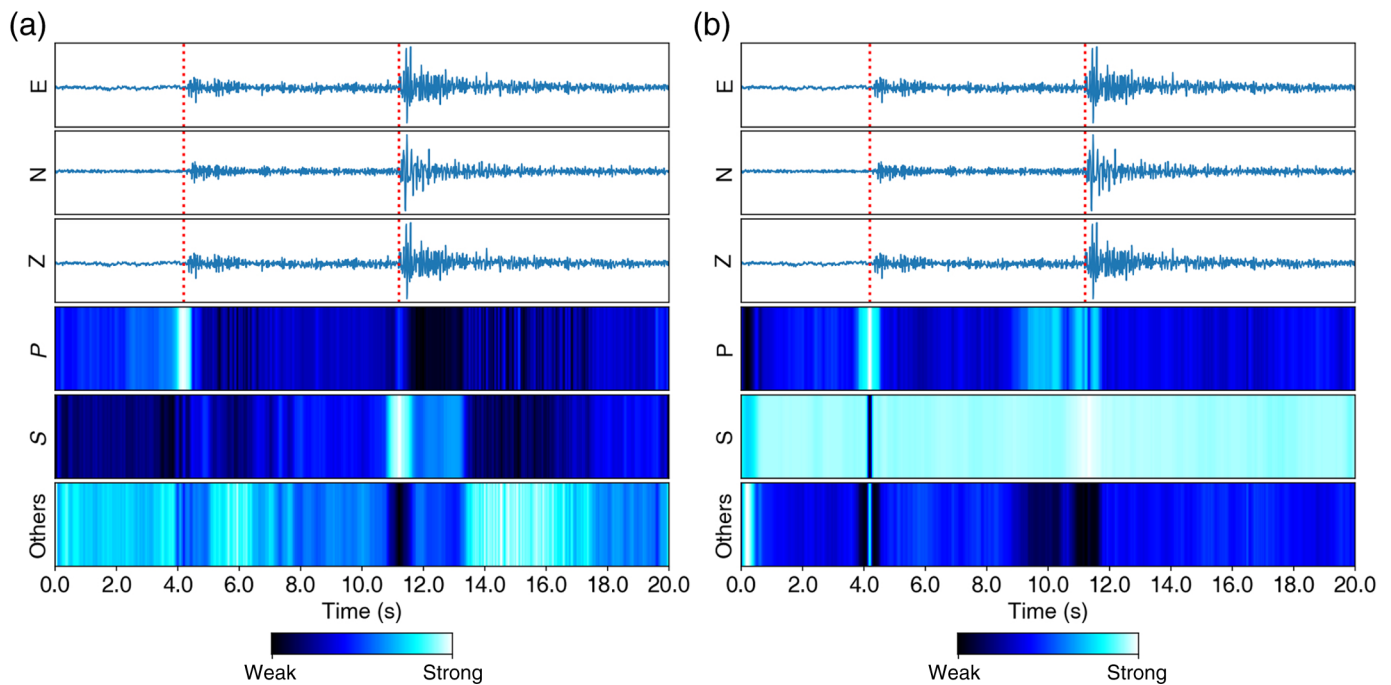


Figure 9. Examples of Gradient-weighted Class activation Maps of (a) the ARR-Net and (b) the U-Net predictions. Colors indicate the strength of confidence, and dotted lines show the manual arrivals. The color version of this figure is available only in the electronic edition.

2017; Golovin *et al.*, 2017; Sculley *et al.*, 2018; Zhu *et al.*, 2020). In seismic phase picking, the data augmentation has been tested, and the results show improvements on both the model performance and generalization. The hyperparameter optimization has not been fully examined in seismic phase-picking methods and is probably a research direction worth considering in future work.

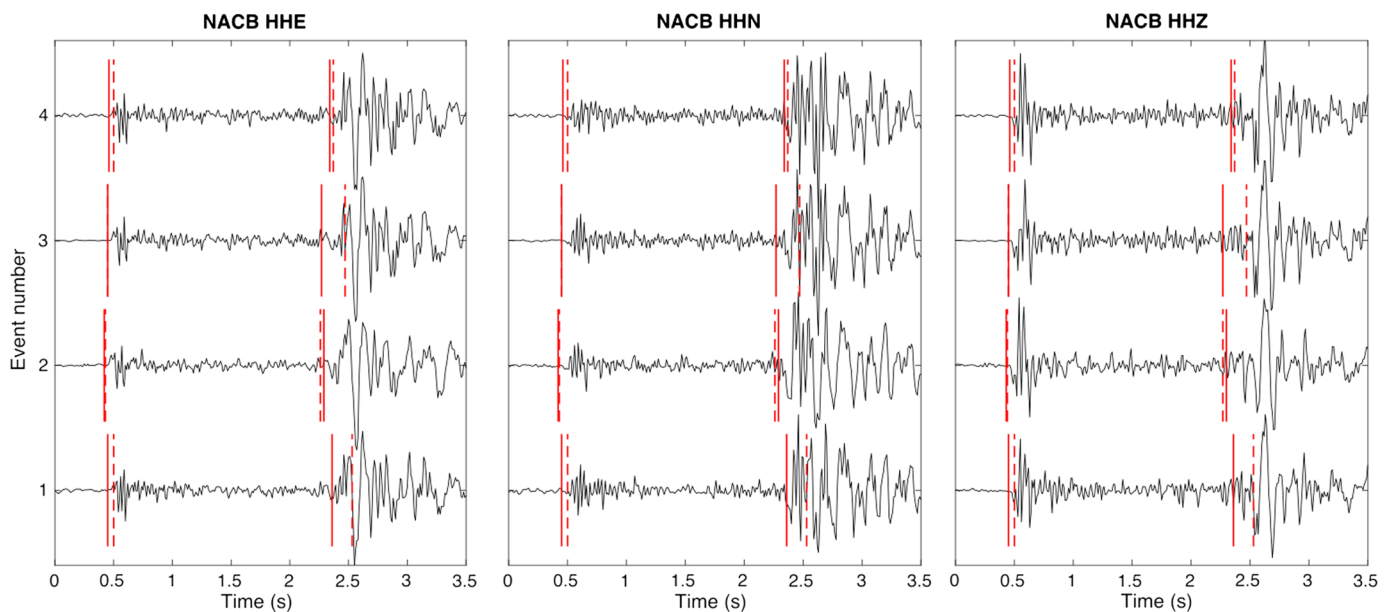


Figure 10. An example shows the differences between the arrivals determined by the ARR-Net model (solid lines) and the manual picks (dash lines) on a set of earthquakes with high-waveform similarities on the three component seismograms (HHE, HHN, HHZ) at

station NACB in BATS. Compared with the manual picks, the arrivals determined by the ARR-Net model are more consistent. The color version of this figure is available only in the electronic edition.

Conclusions

Accurate seismic phase arrivals play an important and essential role in understanding earthquake sources and the Earth interior. Many efforts have been paid to develop reliable and accurate picking algorithms. Compared with the analysis-based phase-picking approaches, the machine-learning-based methods could learn how to pick seismic phases directly from seismograms, without tuning the complex parameter settings. In this study, we tested different U-Net-based machine-learning methods for seismic phase picking and found that the ARRUPhase picker has the best performance. The ARRUPhase picker is built on an efficient U-Net architecture and then included the AGs that increase the weights of seismic phases on seismograms during training and RRCU, strengthening the temporal linkages of features in multiple scales. The included functions show remarkable improvements in both picking *P* and *S* arrivals, especially on the cases with low SNR. In addition, the seismic phase picks provided by ARRUPhase picker also show more consistency at different noise levels (Fig. 10). More accurate seismic phase picks could benefit both the earthquake location and tomographic inversions. The ARRUPhase picker also generalizes well by examining model performance using unseen data (Fig. 1a,b; Table 1) and data from a distinct region (Fig. 1c,d; Table 2). The great model generalization allows the ARRUPhase picker to apply to routine seismic picking tasks in different regions. The ARRUPhase picker can be further combined with earthquake location methods for rapid and accurate earthquake locations (Lee *et al.*, 2020).

Data and Resources

The seismograms of earthquakes occurred in Taiwan used in this study are from the Geophysical Database Management System (GDMS) operated by the Central Weather Bureau (CWB) and the Broadband Array in Taiwan for Seismology (BATS) operated by the Institute of Earth Sciences, Academia Sinica (IES). The Southern California Seismic Network (SCSN) earthquake catalog website is http://service.scedc.caltech.edu/eq-catalogs/date_mag_loc.php (last accessed October 2020). The seismic waveforms in southern California used in this study were obtained from the Southern California Earthquake Data Center (SCEDC) through the Seismogram Transfer Program (STP), and the website of STP is <https://scedc.caltech.edu/data/downloads.html> (last accessed March 2021). The attention gate with recurrent-residual convolution unit (ARRUPhase picker) is available to download from <https://github.com/tso1257771/Attention-Recurrent-Residual-U-Net-for-earthquake-detection> (last accessed February 2021).

Declaration of Competing Interests

The authors declare no competing interests.

Acknowledgments

En-Jui Lee and Wu-Yu Liao are supported by the Ministry of Science and Technology, R.O.C., under Contract MOST 106-2628-M-006-001-MY3. Po Chen acknowledges support from the Nielson Energy

Fellowship provided by the School of Energy Resources, University of Wyoming. The authors thank Southern California Earthquake Data Center (SCEDC) for providing seismic recordings for this study. This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, Grant Number 1725729, as well as the University of Illinois at Urbana-Champaign. The authors also thank National Center for High-performance Computing (NCHC) in Taiwan, for providing computational and storage resources.

References

- Adavanne, S., A. Politis, J. Nikunen, and T. Virtanen (2018). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, *IEEE J. Sel. Top. Signal Process.* **13**, no. 1, 34–48.
- Aki, K., and P. G. Richards (2002). *Quantitative Seismology*, University Science Books, Sausalito, California.
- Allen, R. (1982). Automatic phase pickers: Their present use and future prospects, *Bull. Seismol. Soc. Am.* **72**, no. 6B, S225–S242.
- Alom, M. Z., M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari (2018). Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation, arXiv preprint, available at <https://arxiv.org/abs/1802.06955> (last accessed March 2021).
- Baer, M., and U. Kradolfer (1987). An automatic phase picker for local and teleseismic events, *Bull. Seismol. Soc. Am.* **77**, no. 4, 1437–1445.
- Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural machine translation by jointly learning to align and translate, arXiv preprint, available at <https://arxiv.org/abs/1409.0473> (last accessed March 2021).
- Baillard, C., W. C. Crawford, V. Ballu, C. Hibert, and A. Mangeney (2014). An automatic kurtosis-based *P*- and *S*-phase picker designed for local seismic networks, *Bull. Seismol. Soc. Am.* **104**, no. 1, 394–409.
- Boschi, L., L. Delcor, J.-L. L. Carrou, C. Fritz, A. Paté, and B. Holtzman (2017). On the perception of audified seismograms, *Seismol. Res. Lett.* **88**, no. 5, 1279–1289.
- Chan, W., N. Jaitly, Q. Le, and O. Vinyals (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *2016 IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 4960–4964.
- Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, no. 4, 834–848.
- Chen, P., and E.-J. Lee (2015). *Full-3D Seismic Waveform Inversion: Theory, Software and Practice*, Springer, New York, New York.
- Chetlur, S., C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer (2014). cuDNN: Efficient primitives for deep learning, arXiv preprint, available at <https://arxiv.org/abs/1410.0759> (last accessed March 2021).
- Chiu, C.-C., and C. Raffel (2017). Monotonic chunkwise attention, arXiv preprint, available at <https://arxiv.org/abs/1712.05382> (last accessed March 2021).
- Cichowicz, A. (1993). An automatic *S*-phase picker, *Bull. Seismol. Soc. Am.* **83**, no. 1, 180–189.
- Dai, H., and C. MacBeth (1995a). Automatic picking of seismic arrivals in local earthquake data using an artificial neural network, *Geophys. J. Int.* **120**, no. 3, 758–774.

- Dai, H., and C. MacBeth (1995b). Identifying *P*- and *S*-waves using artificial neural network, *57th EAGE Conf. and Exhibition*, European Association of Geoscientists and Engineers, cp-90-00159.
- Dai, H., and C. MacBeth (1997). The application of back-propagation neural network to automatic picking seismic arrivals from single-component recordings, *J. Geophys. Res.* **102**, no. B7, 15,105–15,113.
- Foggia, P., N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento (2015). Audio surveillance of roads: A system for detecting anomalous sounds, *IEEE Trans. Intell. Transport. Syst.* **17**, no. 1, 279–288.
- Galiana-Merino, J. J., J. L. Rosa-Herranz, and S. Parolai (2008). Seismic *P* phase picking using a Kurtosis-based criterion in the stationary wavelet domain, *IEEE Trans. Geosci. Rem. Sens.* **46**, no. 11, 3815–3826.
- Golovin, D., B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley (2017). Google vizier: A service for black-box optimization, *Proc. of the 23rd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, ACM, 1487–1495.
- Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*, MIT Press, Cambridge, Massachusetts.
- Graves, A., S. Fernández, F. Gomez, and J. Schmidhuber (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, *Proc. of the 23rd International Conf. on Machine Learning*, 369–376.
- Graves, A., A.-R. Mohamed, and G. Hinton (2013). Speech recognition with deep recurrent neural networks, *2013 IEEE International Conf. on Acoustics, Speech and Signal Processing*, IEEE, 6645–6649.
- Guyon, I., and A. Elisseeff (2003). An introduction to variable and feature selection, *J. Mach. Learn. Res.* **3**, 1157–1182.
- He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep residual learning for image recognition, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 770–778.
- Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Process. Mag.* **29**, no. 6, 82–97.
- Holtzman, B., J. Candler, M. Turk, and D. Peter (2013). Seismic sound lab: Sights, sounds and perception of the earth as an acoustic space, *International Symposium on Computer Music Multidisciplinary Research*, Springer, 161–174.
- Iyer, H., and K. Hirahara (1993). *Seismic Tomography: Theory and Practice*, Springer Science & Business Media, London, United Kingdom.
- Jaitly, N., Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, and S. Bengio (2016). An online sequence-to-sequence model using partial conditioning, in *Advances in Neural Information Processing Systems*, 5067–5075.
- Jurkevics, A. (1988). Polarization analysis of three-component array data, *Bull. Seismol. Soc. Am.* **78**, no. 5, 1725–1743.
- Kong, Q., D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft (2019). Machine learning in seismology: Turning data into insights, *Seismol. Res. Lett.* **90**, no. 1, 3–14.
- Krischer, L., T. Megies, R. Barsch, M. Beyreuther, T. Lecocq, C. Caudron, and J. Wassermann (2015). ObsPy: A bridge for seismology into the scientific Python ecosystem, *Comput. Sci. Discov.* **8**, no. 1, 014003.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, 1097–1105.
- Lee, E.-J., and P. Chen (2013). Automating seismic waveform analysis for full 3-D waveform inversions, *Geophys. J. Int.* **194**, no. 1, 572–589.
- Lee, E.-J., P. Chen, and T. H. Jordan (2014). Testing waveform predictions of 3D velocity models against two recent Los Angeles earthquakes, *Seismol. Res. Lett.* **85**, no. 6, 1275–1284.
- Lee, E.-J., P. Chen, T. H. Jordan, P. B. Maechling, M. A. Denolle, and G. C. Beroza (2014). Full-3-D tomography for crustal structure in southern California based on the scattering-integral and the adjoint-wavefield methods, *J. Geophys. Res.* **119**, no. 8, 6421–6451.
- Lee, E.-J., W.-Y. Liao, G.-W. Lin, P. Chen, D. Mu, and C.-W. Lin (2019). Towards automated real-time detection and location of large-scale landslides through seismic waveform back projection, *Geofluids* **2019**, doi: [10.1155/2019/1426019](https://doi.org/10.1155/2019/1426019).
- Lee, E.-J., W.-Y. Liao, D. Mu, W. Wang, and P. Chen (2020). GPU-accelerated automatic microseismic monitoring algorithm (GAMMA) and its application to the 2019 Ridgecrest earthquake sequence, *Seismol. Res. Lett.* **91**, no. 4, 2062–2074.
- Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization, *J. Mach. Learn. Res.* **18**, no. 1, 6765–6816.
- Liang, M., and X. Hu (2015). Recurrent convolutional neural network for object recognition, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 3367–3375.
- Lomax, A., C. Satriano, and M. Vassallo (2012). Automatic picker developments and optimization: FilterPicker—A robust, broadband picker for real-time seismic monitoring and earthquake early warning, *Seismol. Res. Lett.* **83**, no. 3, 531–540.
- Mousavi, S. M., W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nat. Comm.* **11**, Article Number 3952.
- Mousavi, S. M., Y. Sheng, W. Zhu, and G. C. Beroza (2019). Stanford Earthquake Dataset (STEAD): A global data set of seismic signals for AI, *IEEE Access* **7**, 179,464–179,476.
- Mu, D., E.-J. Lee, and P. Chen (2017). Rapid earthquake detection through GPU-based template matching, *Comput. Geosci.* **109**, 305–314, doi: [10.1016/j.cageo.2017.09.009](https://doi.org/10.1016/j.cageo.2017.09.009).
- Nippres, S., A. Rietbrock, and A. Heath (2010). Optimized automatic pickers: Application to the ANCORP data set, *Geophys. J. Int.* **181**, no. 2, 911–925.
- Paté, A., L. Boschi, D. Dubois, J.-L. Le Carrou, and B. Holtzman (2017). Auditory display of seismic data: On the use of experts' categorizations and verbal descriptions as heuristics for geoscience, *J. Acoust. Soc. Am.* **141**, no. 3, 2143–2162.
- Paté, A., L. Boschi, J.-L. Le Carrou, and B. Holtzman (2016). Categorization of seismic sources by auditory display: A blind test, *Int. J. Hum. Comput. Stud.* **85**, 57–67.
- Perol, T., M. Gharbi, and M. Denolle (2018). Convolutional neural network for earthquake detection and location, *Sci. Adv.* **4**, no. 2, e1700578.
- Ronneberger, O., P. Fischer, and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation, *International Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, 234–241.
- Ross, Z. E., and Y. Ben-Zion (2014). Automatic picking of direct *P*, *S* seismic phases and fault zone head waves, *Geophys. J. Int.* **199**, no. 1, 368–381.

Ross, Z. E., M.-A. Meier, and E. Hauksson (2018). *P* wave arrival picking and first-motion polarity determination with deep learning, *J. Geophys. Res.* **123**, no. 6, 5120–5129.

Ross, Z. E., M.-A. Meier, E. Hauksson, and T. H. Heaton (2018). Generalized seismic phase detection with deep learning, *Bull. Seismol. Soc. Am.* **108**, no. 5A, 2894–2901.

Ross, Z. E., D. T. Trugman, E. Hauksson, and P. M. Shearer (2019). Searching for hidden earthquakes in Southern California, *Science* **364**, no. 6442, 767–771, doi: [10.1126/science.aaw6888](https://doi.org/10.1126/science.aaw6888).

Saragiotis, C. D., L. J. Hadjileontiadis, and S. M. Panas (2002). PAI-S/K: A robust automatic seismic *P* phase arrival identification scheme, *IEEE Trans. Geosci. Rem. Sens.* **40**, no. 6, 1395–1404.

Schlemper, J., O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert (2019). Attention gated networks: Learning to leverage salient regions in medical images, *Med. Image Anal.* **53**, 197–207.

Sculley, D., J. Snoek, A. Rahimi, and A. Wiltschko (2018). Winner's curse? On pace, progress, and empirical rigor, *ICLR Workshop*, 4 pp.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra (2017). Grad-Cam: Visual explanations from deep networks via gradient-based localization, *Proc. of the IEEE International Conf. on Computer Vision*, 618–626.

Shearer, P. M. (1997). Improving local earthquake locations using the L1 norm and waveform cross correlation: Application to the Whittier Narrows, California, aftershock sequence, *J. Geophys. Res.* **102**, 8269–8283.

Shin, T.-C., C.-H. Chang, H.-C. Pu, H.-W. Lin, and P.-L. Leu (2013). The geophysical database management system in Taiwan, *Terr. Atmos. Ocean. Sci.* **24**, no. 1, 11.

Shin, T.-C., Y.-B. Tsai, and Y.-M. Wu (1996). Rapid response of large earthquakes in Taiwan using a real-time telemetered network of digital accelerographs, *Proc. 11th World Conf. on Earthquake Engineering*, Paper Number 2137.

Sleeman, R., and T. Van Eck (1999). Robust automatic *P*-phase picking: An on-line implementation in the analysis of broadband seismogram recordings, *Phys. Earth Planet. In.* **113**, nos. 1/4, 265–275.

Southern California Earthquake Data Center (SCEDC) (2013). Southern California Earthquake Data Center, available at <https://scedc.caltech.edu/> (last accessed March 2021).

Storchak, D. A., J. Schweitzer, and P. Bormann (2003). The IASPEI standard seismic phase list, *Seismol. Res. Lett.* **74**, no. 6, 761–772.

Storchak, D. A., J. Schweitzer, and P. Bormann (2011). Seismic phase names: IASPEI standard, in *Encyclopedia of Solid Earth Geophysics*, Springer, 1162–1173.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016). Rethinking the inception architecture for computer vision, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2818–2826.

Vincent, L., and P. Soille (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations, *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, no. 6, 583–598.

Waldhauser, F., and W. L. Ellsworth (2003). A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, California, *Bull. Seismol. Soc. Am.* **90**, no. 6, 1353–1368.

Wang, J., and T.-L. Teng (1995). Artificial neural network-based seismic detector, *Bull. Seismol. Soc. Am.* **85**, no. 1, 308–319.

Wang, J., and T.-L. Teng (1997). Identification and picking of *S* phase using an artificial neural network, *Bull. Seismol. Soc. Am.* **87**, no. 5, 1140–1149.

Yu, S.-B., H.-Y. Chen, and L.-C. Kuo (1997). Velocity field of GPS stations in the Taiwan area, *Tectonophysics* **274**, no. 1, 41–59.

Zhang, H., and C. H. Thurber (2003). Double-difference tomography: The method and its application to the Hayward fault, California, *Bull. Seismol. Soc. Am.* **93**, 1875–1889.

Zhang, H., C. H. Thurber, and C. Rowe (2003). Automatic *P*-wave arrival detection and picking with multiscale wavelet analysis for single-component recordings, *Bull. Seismol. Soc. Am.* **93**, no. 5, 1904–1912.

Zhao, Y., and K. Takano (1999). An artificial neural network approach for broadband seismic phase picking, *Bull. Seismol. Soc. Am.* **89**, no. 3, 670–680.

Zhu, L., Z. Peng, J. McClellan, C. Li, D. Yao, Z. Li, and L. Fang (2019). Deep learning for seismic phase detection and picking in the aftershock zone of 2008 M_w 7.9 Wenchuan earthquake, *Phys. Earth Planet. In.* **293**, 106261.

Zhu, W., and G. C. Beroza (2019). PhaseNet: A deep-neural-network-based seismic arrival-time picking method, *Geophys. J. Int.* **216**, no. 1, 261–273.

Zhu, W., S. M. Mousavi, and G. C. Beroza (2020). Seismic signal augmentation to improve generalization of deep neural networks, in *Advances in Geophysics*, Vol. 61, Elsevier, 151–177.

Appendix

Here, we adapt the 2D formulation for recurrent convolutional layer, in particular equations (1)–(4) in Liang and Hu (2015), to the 1D seismic data used in our study. To aid the understanding of the formulation, we refer to the left panel of figure 3 in Liang and Hu (2015) for the unfolded representation of the recurrent connections. For the 1D seismic data used in our study, the rectangles used for representing 2D images in figure 3 of Liang and Hu (2015) can be replaced with 1D line segments, and the two spatial axes can be replaced with one time axis. To avoid confusion with the discrete time sampling of the seismic data, we introduce the recurrent iteration index $q = 0, \dots, Q - 1$, in which $Q \geq 1$, to replace the notion of "time step" t used in figure 3 of Liang and Hu (2015).

For the l th recurrent-residual convolutional layer, we define the net input to the i th unit on the k th channel of the feature map at the $(q + 1)$ th recurrent iteration as

$$z_{kl}^{(q+1)}(i) = \sum_{j=i-L_F}^{i+L_F} [\mathbf{W}_{kl}^F(j-i+L_F)]^T \mathbf{x}_l(j) + \sum_{j=i-L_R}^{i+L_R} W_{kl}^R(j-i+L_R) h_{kl}^{(q)}(j) + b_{kl}, \quad (\text{A1})$$

in which i and j are temporal indices, $\mathbf{x}_l(j) \in \mathbb{R}^{K_l}$ is the feedforward input from the previous layer with K_l channels, and $\mathbf{x}_0 = \mathbf{X}$, $h_{kl}^{(q)}$ is the recurrent input from the state \mathbf{H} at the q th iteration, \mathbf{W}_{kl}^F and W_{kl}^R are the shared weights corresponding to the feedforward and recurrent inputs, respectively, and b_{kl} is

the bias. In this equation, the summation over the input temporal index j corresponds to the convolution operation. The number of rows (temporal dimension) and columns of the shared weights \mathbf{W}_{kl}^F is $2L_F + 1$ and K_l , respectively. As j ranges from $i - L_F$ to $i + L_F$, the index $j - i + L_F$ into \mathbf{W}_{kl}^F ranges from 0 to $2L_F$. The number of rows (temporal dimension) and columns of the weights \mathbf{W}_{kl}^R is $2L_R + 1$ and 1, respectively. As j ranges from $i - L_R$ to $i + L_R$, the index $j - i + L_R$ into \mathbf{W}_{kl}^R ranges from 0 to $2L_R$.

For a given output temporal index i , the rows $i - L_F$ to $i + L_F$ in \mathbf{x}_l and the rows $i - L_R$ to $i + L_R$ in $\mathbf{h}_{kl}^{(q)}$ are selected. The widths of the receptive fields of the i th unit for the feedforward and recurrent connections are $2L_F + 1$ and $2L_R + 1$, respectively. Increasing the recurrent iteration index q by 1 will extend the width of the recurrent receptive field of each unit in the k th feature map of layer l by $2L_R + 1$, and the depth of the longest path will also increase by 1. However, because of weight sharing, the number of parameters will not increase with q .

The state at iteration $q = 0$ is $\mathbf{h}_{kl}^{(0)} = 0$ and the net input $\mathbf{z}_{kl}^{(1)}$ is identical to that in the standard CNN. The state at the $(q + 1)$ th iteration is computed from $\mathbf{z}_{kl}^{(q+1)}$, that is, $\mathbf{h}_{kl}^{(q+1)}(j) = g\{f[\mathbf{z}_{kl}^{(q+1)}(j)]\}$, in which $f[\cdot]$ is the rectified linear (ReLU) activation function, and $g\{\cdot\}$ is the local response normalization function (Krizhevsky *et al.*, 2012; Liang and Hu, 2015), for preventing the state from exploding.

To ease the training of our network, which can be deeper than the one without the recurrent iterations, we make use of residual connections (He *et al.*, 2016; Alom *et al.*, 2018), that is, the net input is passed into the activation function, to produce the residual output and the net output of layer l is $\mathbf{x}_{l+1} = \mathbf{x}_l + \mathcal{F}(\mathbf{x}_l)$, in which the function $\mathcal{F}(\cdot)$ represents all transformations within the layer that produce the residual output from the input \mathbf{x}_l .

To compute the grid-based attention coefficients $\alpha_l(i)$ for the input feature map $\mathbf{x}_l(i)$, we need the feature map at a

coarser level $\mathbf{x}_g(j) \in \mathbb{R}^{K_g}$ as the gating signal. The gating signal at the same position as the input feature map $\mathbf{x}_g(i)$ can be obtained through temporal linear interpolation (Schlemper *et al.*, 2019). In our implementation, we adopt the additive attention (Bahdanau *et al.*, 2014; Schlemper *et al.*, 2019)

$$\alpha_l(i) = \sigma\{\Psi^T f[\mathbf{y}_l^{(0)}(i) + \mathbf{y}_l^{(Q)}(i)] + b_\Psi\}, \quad (\text{A2})$$

in which

$$\mathbf{y}_l^{(0)}(i) = \mathbf{W}_l^T \mathbf{x}_l^{(0)}(i) + \mathbf{W}_g^T \mathbf{x}_g^{(0)}(i) + \mathbf{b}_{lg} \quad (\text{A3})$$

is the contribution from the feedforward connections and is identical to previous studies (Schlemper *et al.*, 2019) and

$$\mathbf{y}_l^{(Q)}(i) = \mathbf{V}_l^T \mathbf{x}_l^{(Q)}(i) + \mathbf{V}_g^T \mathbf{x}_g^{(Q)}(i) \quad (\text{A4})$$

is the contribution from the recurrent connections. Here, \mathbf{W}_l , \mathbf{V}_l and \mathbf{W}_g , \mathbf{V}_g are trainable weight matrices that transform the corresponding feature vectors into vectors of the same dimension as the bias vector \mathbf{b}_{lg} . In our current implementation, we choose $\mathbf{V}_l = \mathbf{W}_l$ and $\mathbf{V}_g = \mathbf{W}_g$, the ReLU activation function $f[\cdot]$ produces a vector of the same dimension as the trainable parameter vector Ψ , b_Ψ is a scalar bias, and $\sigma\{\cdot\}$ is the sigmoid activation function. Once the attention coefficient $\alpha_l(i)$ is computed for every temporal index i , using equation (A2), we can obtain the attention-gated feature map through the element-wise scalar multiplication, that is, $\mathbf{x}_l(i) = \alpha_l(i)\mathbf{x}_l(i)$. The attention-gated feature map \mathbf{x}_l is then concatenated into the corresponding level of the decoder through the skip connection in the attention U-Net architecture (Schlemper *et al.*, 2019; Fig. 3a).

Manuscript received 13 October 2020

Published online 24 March 2021